THE UNIVERSITY *of* EDINBURGH

Genomic catastrophes: Complex Structural Variants in High Grade Serous Ovarian Cancer

THE UNIVERSITY *of* EDINBURGH

Stuart L. Brown

Doctor of Philosophy
The University of Edinburgh
2024

Declaration

This thesis presents my own work, wherever the contributions of others were involved this is clearly indicated. It has not been submitted for any other degree or professional qualification. – Stuart L. Brown

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisors, Professor Colin Semple and Dr Ailith Ewing for their invaluable guidance, encouragement, and support throughout my doctoral journey. Without their expert advice, I would not have been able to complete this thesis and it has been a privilege to learn from them.

I would like to acknowledge the unwavering support and encouragement of my parents, Douglas Brown and Denise Brown, who have been my constant source of inspiration, humour, and motivation during this PhD and the years leading up to it. Without their superhuman and consistent efforts, I would not have been able to undertake this PhD. It is an honour to have them as role models, supporters, and parents.

Finally, I would like to express my gratitude to my girlfriend, Divjot Kaur, for her steadfast support and encouragement throughout my academic journey. Her love, patience, and understanding have been essential in helping me navigate the challenges of pursuing a PhD. I am grateful to have her by my side as my partner and inspiration.

This may have been the most challenging page to write in my thesis. If I could adequately convey my gratitude and appreciation to the people mentioned above, I would be a poet.

Abstract

High-grade serous ovarian cancer (HGSOC) is the most common and lethal type of ovarian cancer. HGSOC accounts for 70-80% of deaths from all forms of ovarian cancer, roughly 98000-110000 deaths worldwide each year according to WHO estimates. It is characterized by ubiquitous mutations in the tumour suppressor gene TP53. Roughly half of HGSOC tumours also have mutations in genes involved in the homologous recombination repair pathway, primarily BRCA1 and BRCA2, leading to homologous recombination deficiency (HRD). Cancers with HRD must rely on alternative DNA repair pathways to repair breaks in their DNA. The PARP protein is key to the base excision repair/single-strand break repair pathway, and recent HGSOC drugs inhibit the PARP protein to make HRD tumours unable to efficiently repair genomic breaks and leading to apoptosis. However, these cancers often gain resistance to PARP inhibitors by regaining the ability to use the homologous recombination DNA repair pathways, which renders PARP inhibitors ineffective. On average only 45% of patients diagnosed with HGSOC survive for more than 5 years after diagnosis. HGSOC is highly genomically rearranged and contains hundreds to thousands of structural variants, though their impacts on tumour function and evolution are poorly understood. Structural variants usually accumulate over many cell cycles, but it has recently become clear that large numbers can be acquired in a single cell cycle in complex patterns called complex structural variants (cSV). It is known that cSVs are catastrophic genomic events that can generate hundreds of structural changes across the genome, and at least some cSV types have been reported in HGSOC. The changes induced by cSVs can increase or decrease the copy number of genes, change their genomic location and orientation, or disrupt the gene entirely. The increased use of whole genome sequencing in large cancer cohorts has allowed an increasing number of cSV types to be identified. However, investigations into cSVs often focus on a single cSV type, frequently using differing computational criteria for the identification of cSVs. This means that the relationships between cSVs remains understudied, and interpretability of the impacts of cSVs between studies is challenging.

In this work, eight cSV types were identified across the 324 whole-genome-sequenced HGSOC samples using previously published criteria: chromothripsis, chromoplexy, breakage fusion bridges, ecDNA, pyrgo, rigma, tymphonas and seismic amplification. The prevalence, distribution, and impact on survival of each cSV type was assessed, and their relationships with other genomic features, such as HRD and whole genome duplication (WGD) were investigated. By studying eight cSV types together across a uniformly processed well annotated cohort, many novel insights into HGSOC structural evolution were gained. I have shown that there are two main routes to genomic diversity in the HGSOC cohort. One of these routes involves HRD while the other route involves WGD and cSVs, and it appears that WGD can buffer the deleterious effects of cSV. I also found that the presence of cSVs was not associated with significant variation in survival rates, but there was a trend that patients with tumours containing more severe cSV events had better survival rates compared to those with less severe events of the same type. This was observed for chromothripsis, a particularly disruptive cSV often described as 'chromosome shattering'. My results also revealed a novel hotspot for multiple cSV types on chromosome 19 which covers the known HGSOC oncogene CCNE1, which confers worse survival on patients when amplified. I have shown that the amplification of CCNE1 by breakage fusion bridges results in higher copy numbers than by simple amplification. Overall, this thesis investigates the prevalence, co-occurrence, distribution, and impact of multiple cSV types on survival in HGSOC and provides novel insights into the genomic diversity of HGSOC.

Lay Summary

The rapid generation of multiple large changes within a cancer cell's DNA are called complex structural variants (cSV). These changes can result in genes changing in number, location, orientation or even becoming non-functional. Whole genome sequencing allows for all the DNA in the cell to be investigated. As larger cohorts of cancer samples have been whole genome sequenced, an increasing number of different cSV patterns have been identified. However, the investigation of cSV is often limited to a single pattern of cSV.

By utilising the largest cohort of whole genome sequenced high grade serous ovarian cancer and identifying eight previously published patterns of cSV. The breadth of cSV patterns identified found that there were two main ways these changes occurred: some samples had defects in DNA repair, while others had duplicated their entire genome before experiencing cSV. Interestingly, while cSV did not generally affect patient survival rates, the more severe cases of cSV tended to be associated with better survival.

Table of Contents

Table of Figure

Table of Tables

List Of Acronyms

Acronym	Description
AOCS	Australian Ovarian Cancer Study
BCCA	British Columbia Cancer Agency
BFB	Breakage-fusion-bridge
CN	Copy Number
CNV	Copy Number Variant
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
cSV	Complex Structural Variant
ecDNA	Extrachromosomal Circular DNA
HGSOC	High-Grade Serous Ovarian Cancer
JaBbA	Junction Balance Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LRS	Long Read Sequencing
MDA	University of Texas MD Anderson Cancer Center
PCR	Polymerase Chain Reaction
SA	Seismic Amplification
SHGSOC	Scottish High Grade Serous Ovarian Cancer
SNV	Single Nucleotide Variant
SRS	Short Read Sequencing
SV	Structural Variant
TCGA	The Cancer Genome Atlas
UMP	Underlying Mutational Process
WGD	Whole Genome Duplication
WGS	Whole Genome Sequencing
WHO	World Health Organization

“He spent a year in silence just to better understand the sound of a whisper.”
- Chaucer (A Knight’s Tale)

Chapter 1: Introduction

Cancer is commonly referred to as a genetic disease that occurs due to the accumulation of mutations over numerous cell cycles (Hanahan and Weinberg 2000). These mutations cause healthy cells to slowly transform into cancerous ones through point mutations and structural rearrangements (Hanahan and Weinberg 2000). Recently, bursts of mutations occurring within a single cell cycle and generating hundreds to thousands of mutations have been reported and classified as complex structural variant types which offer an intriguing and understudied new area of cancer genomics (Baca et al. 2013; Stephens et al. 2011).

Chromosomal rearrangements were first linked to inherited diseases in 1960 based upon cytogenetic chromosome fixing and staining (Moorhead et al. 1960; Durmaz et al. 2015). Since then, array based techniques have enabled sections of the genome to be looked at in greater detail allowing new investigations of cancer genetics. Although arrays continue to play an important role in certain investigations of parts of the genome, these methods are inappropriate for the study of large genome rearrangements, where whole genome sequencing (WGS) based methods have become dominant (Rezaee et al. 2022; Hoheisel 2006; Gresham, Dunham, and Botstein 2008; Kinsella and Bafna 2012). The Human Genome Project published the first draft human genome reference sequence in 2001 (Lander et al. 2001). Advances in molecular techniques and high throughput sequencing technology have progressively reduced the cost of WGS for an individual, from a hundred million dollars in 2001 to less than a thousand dollars (Church and Gilbert 1984; Metzker 2009; Schwarze et al. 2018; Singh 2018).

With the decreasing cost barriers to WGS, this more comprehensive source of information was used to investigate disease genetics, ranging from rare undiagnosed diseases to cancer (Mittelman and Wilson 2013; Souche et al. 2022). Studies in cancer genetics using WGS have included the study of immortalised cell lines derived from

patients' cancer cells, primary samples taken from patients with a single tumour type and pan-cancer cohorts of thousands of patients with different tumour types (Weinstein et al. 2013; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisiation/Ins et al. 2020; 100,000 Genomes Project Pilot Investigators et al. 2021). These studies have allowed changes in tumour genomes to be directly observed and classified into three main types: single nucleotide variants (SNVs) such as indels and point mutations, structural variants (SVs) and copy number variants (CNVs).

SNVs refer to alterations of one base pair that can significantly affect the survival rates in numerous cancer types, especially when they occur within coding exons of specific genes. This area of research is highly active and extensively studied, providing valuable insights into cancer biology. (Goh, Yao, and Smith 1995; Hart et al. 2015; He et al. 2014). SVs and CNVs are relatively understudied in comparison to SNVs, particularly in terms of their functional impact on genes (Mahmoud et al. 2019). Encompassing at least 50 bp, an SV can describe changes in the orientation, location or DNA content of parts of a chromosome (Mahmoud et al. 2019). These alterations can affect gene structures, resulting in disruptions of genes in various tumour types (Mahmoud et al. 2019). CNVs are large gains and losses of DNA, often spanning many megabases of sequence, and have been shown to impact cancer progression and response to treatment (Steele et al. 2022). Patterns of SVs and CNVs occurring together more often than would be expected by chance have been identified as complex structural variants (cSVs) and appear to have important roles in tumour biology (Mahmoud et al. 2019).

Complutational models of complex structural variants suggest mutational mechanisms where tens to thousands of structural variants occur simultaneously within a single cell cycle, in a sudden rather than progressive acquisition of genomic changes (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). The occurrence, underlying mechanism and impact of cSVs are understudied. But they may play particularly important roles in highly rearranged tumour types such as high grade serous ovarian cancer (HGSOC) (Mittelman and Wilson 2013) (The ICGC/TCGA Pan- Cancer Analysis of Whole Genomes Consortium (organisiation/Ins et al. 2020).

The main objective of this thesis is to determine the roles of cSVs in a HGSOC. To achieve

this, the latest analysis approaches will be used to examine the occurrence, patterns, and functional impacts of these cSVs. The research will utilize the largest uniformly processed WGS HGSOC dataset to investigate the cSVs thoroughly.

High Grade Serous Ovarian Cancer

HGSOC accounts for 70-80% of deaths from all forms of ovarian cancer, roughly 98000-110000 deaths each year according to WHO estimates, and is a leading cause of cancer death in women (Lisio et al. 2019; Jemal et al. 2010). Despite its name HGSOC is now widely accepted to originate from the epithelium of the fallopian tubes (Labidi-Galy et al. 2017; Bergsten, Burdette, and Dean 2020; Karst and Drapkin 2010; Przybycin et al. 2010; Crum 2009). The metastasis of HGSOC occurs via direct spread to adjacent tissues unlike many epithelial cancers that spread via the vascular or lymphatic systems. HGSOC mainly spreads to organs or structures within the peritoneal cavity, particularly the large adipose tissue layer on the surface of intraperitoneal organs called the omentum which is the target of metastasis in 80% of HGSOC patients (Lisio et al. 2019).

The symptoms associated with HGSOC are often diverse and non-specific and, there are currently no effective screening strategies for HGSOC resulting in 75-80% of cases being diagnosed at a late stage (Lisio et al. 2019; Menon et al. 2015; Jacobs et al. 2016). Studies that have reported early detection failed to improve patient survival (Lisio et al., 2019; Menon et al., 2015, 2021; Jacobs et al., 2016). Standard treatments involve surgery to remove as much tumour mass as possible, preceded and or followed by platinum-taxane chemotherapy (Bell D. et al. 2012). However, Some patitients do not respond to chemotherapy and of those which do within 6 months, 25% of patients will have developed a platinum-resistant cancer and there is an overall 5 year survival probability of 31% (Bell D. et al. 2012).

HGSOC Genomic Landscape and DNA Repair Deficiencies

HGSOC is a type 2 ovarian cancer as defined by Kurman and Shih in 2004, characterised by TP53 mutations (in upwards of 96% of samples) and genomic instability due to

defects in pathways contributing to DNA repair (Lisio et al. 2019; I.-M. Shih and Kurman 2004). However, other than TP53 there are few recurrent SNV mutations Table 1 (Bell D. et al. 2012; Lisio et al. 2019; Ahmed et al., 2010).

Gene	Prevalence
TP53	>96%
BRCA1	12.5%
BRCA2	11.5%
CSMD3	6%
NF1	4%
CDK12	3%
FAT3	6%
GABRA6	2%
RB1	2%

Table 1 Recurrent mutations in genes found in HGSOC

Frequency of common single nucleotide variants in genes in HGSOC based upon whole-exome sequencing data (Bell D. et al. 2012; Lisio et al. 2019).

The homologous recombination repair pathway is involved in the repair of double strand breaks. (X. Li and Heyer 2008). When this system of DNA repair is impeded, genomic instabliity increases, and this disruption is called homologous recombination repair deficiency (HRD) and plays an important role in HGSOC (Takaya et al. 2020).Two key genes in this pathway are BRCA1 and BRCA2 and the presence of mutations in these genes can be used to identify HRD (Takaya et al. 2020). The synthetic lethality relationship between BRAC1 and 2 and PARP means that cells that are HRD are sensitized to drugs targeting PARP1 and PARP2 genes. These drugs trap the PARP repair proteins resulting in DNA double strand breaks and ultimately cell death for cells that are sensitized to PARP inhibiters by HRD (Takaya et al. 2020). As the non-cancers cells within the patient will have functional homologous recombination repair pathways although the PARP inhibiter have a impact is systemic there lethality is selective to the cancer cells that are HRD (Takaya et al. 2020).

PARP inhibitor drugs are now often used in the treatment of HGSOC cases with HRD (Takaya et al. 2020). Previously, the mutations identified to be impacting BRCA1 and BRCA2 were SNVs. However, large multi-megabase SVs have also been identified

affecting these genes in HGSOC and resulting in HRD (Ewing et al. 2020).

Much of the genomic data for HGSOC was produced from three landmark studies, two pan-cancer studies that have profiled and analysed HGSOC in addition to other cancer types, The Cancer Genome Atlas (TCGA) in 2013 and the Pan-Cancer Analysis of Whole Genomes (PCAWG) in 2020 (Weinstein et al. 2013; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisiation/Ins et al. 2020), plus another study focussed on chemoresistant and relapsed cases of HGSOC (Patch et al, 2015). The TCGA study was based upon whole-exome sequencing for 316 samples and also covered array based CNV detection, array based promoter methylation, and RNA-seq expression data.

Subsequently, 44 of the samples used in the TCGA study were whole genome sequenced. In addition to identifying the genes found in Table 1, TCGA noted the high level of genomic instability in HGSOC when compared to other tumour types. (Weinstein et al. 2013). The PCAWG study profiled 113 HGSOC samples with WGS and RNA-seq and provided the first comprehensive picture of SVs in this tumour type, also examining, the occurrence of a particular cSV, chromothripis, using a tool called ShatterSeek developed by members of the PCAWG consortium (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisation/Ins et al. 2020; Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). Patch et al (2015) studied HGSOC cases in which many patients aquired resistance to treatment and identified SVs causing common amplification of CCNE1 and inactivation of several tumor suppressor genes (RB1, NF1, RAD51B and PTEN) as well as frequent SNV inactivation of BRCA1 and BRCA2 (Patch et al. 2015). However, the amplification of CCNE1 has also been reported not to correlate with chemoresistance of HGSOC so its amplification may be a more general feature of HGSOC (Sapoznik et al. 2017). This thesis will further investigate SVs and cSVs in HGSOC.

Genomic and expression profiling of a combined HGSOC cohort

In advance of the work presented in this thesis, a combined cohort was constructed by the Gourley, Ewing and Semple labs and consists of HGSOC samples from the Australian Ovarian Cancer Study (AOCS) (Patch et al. 2015), The Cancer Genome Atlas (TCGA) (Bell

D. et al. 2012), the British Columbia Cancer Agency (BCCA) (Wang et al. 2017), The University of Texas MD Anderson Cancer Center (MDA) (Lee et al, 2020), and the Scottish High Grade Serous Ovarian Cancer (SHGSOC) consortium. This cohort (n = 324; Figure 1) provides the largest uniformly processed HGSOC dataset with deep tumour

WGS coverage (75X or greater), matched normal blood (30X) WGS for all samples, and RNA-seq expression data for most samples.

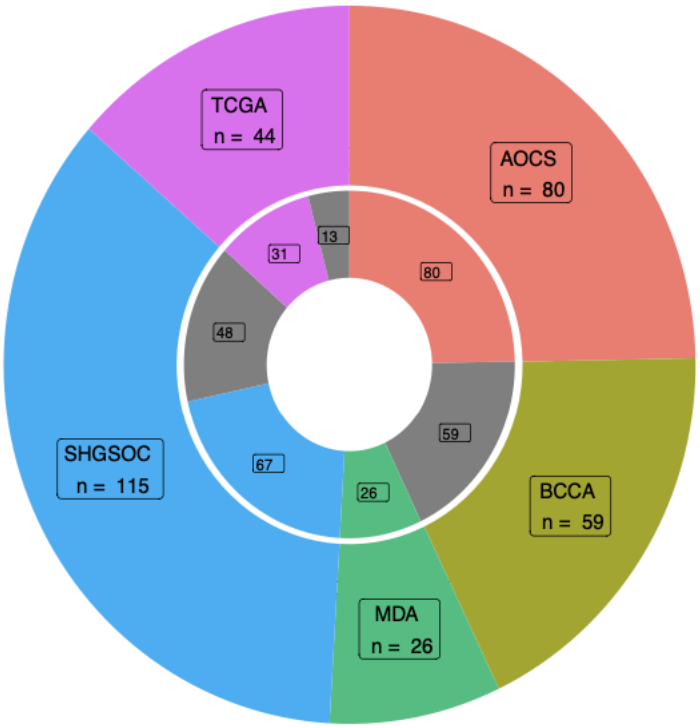


Figure 1 Overview of the HGSOC combined cohort

Sub-cohorts which make up the combined cohort are from the Australian Ovarian Cancer Study (AOCS), The Cancer Genome Atlas (TCGA), the British Columbia Cancer Agency (BCCA), The University of Texas MD Anderson Cancer Center (MDA) and, the Scottish High Grade Serous Ovarian Cancer (SHGSOC) consortium. The outer circle indicates the numbers of WGS samples included, while the inner circle shows the number of samples from each sub-cohort with RNA-seq data. The gray sections represent samples without gene expression data within each sub-cohort.

All tumour samples within the combined cohort were taken during the initial debulking surgery and are pre-treatment with the exeption of the SHGSOC which had neo-adjuvant chemo. In addition to the technical differences that exist between sub-cohorts (such as WGS and RNA-seq coverage), there is also a biological difference between the AOCS sub-cohort and the rest of the combined cohort. The AOCS is comprised of HGSOC samples which was bised towards resistances as they selected patients after progression following platinum based chemotherapy treatment (Patch et

al. 2015), while the other cohorts are comprised of a mixture of samples from patients both responsive and non-responsive to platinum-based chemotherapy treatment. This biological difference can be exploited to investigate the relationship between cSV occurrence and resistance to treatment. This dataset also contains matched RNAseq expression data for 204 (62%) of the WGS tumour samples (Figure 1 Inner Ring) which will be used to investigate the consequences of cSV on gene expression. The combined cohort also has associated clinical data such as age, tumour stage, overall patient survival time after diagnosis and time to relapse. These data can be used to investigate the impact of cSV on patient outcomes.

The combined cohort was uniformly processed by other members of the Semple and Ewing labs to call simple genomic alterations (CNV and SV) and these calls form the basis of the cSV predictions. The importance of CNVs and SVs in tumourigenesis is well established.

Copy Number Variation in Tumorigenesis

Although precise detection of CNVs is an ongoing challenge relying on change of coverage depth this limits the resolution to which CNV can be called, unlike SNV where the exact base pair change is know, CVN start and end cannot be called to an exact base pair. Large scale gains and losses of DNA also known as CNV can be detected via WGS. Changes in coverage depth across large chromosome spans are critical to CNV detection by a variety of algorithms (Coutelier et al. 2022; Talevich et al. 2016; Cameron, Baber, Shale, Papenfuss, et al.
2019). CNV are prevalent across the genome, affecting a larger proportion of genome than SNVs and impacting a range of human phenotypes (Sansregret and Swanton 2017; Shlien and Malkin 2009; M. Zhao et al. 2013; Steele et al. 2022).

Several cancer types have been found to have CNVs, which affect numerous oncogenes and tumour suppressor genes. (Shlien and Malkin 2009; M. Zhao et al. 2013). It has also been reported that an increased number of CNVs in the genome has been associated with higher tumour grades and worse prognosis (Levine and Holland 2018; Glassman 2000). More recently, genome wide patterns of CNV, or CNV signatures, have been reported to have associations with clinically important features of tumour biology such

as HRD and sensitivity to platinum treatment (Shao et al. 2019; Steele et al. 2022). Changes in gene copy number in tumors have also been reported to be highly correlated with increased or decreased gene expression (Shao et al. 2019).

Whole genome duplication (WGD) is an event that can account for significant changes in copy number across the majority of the genome. It involves the doubling of ploidy throughout the entire genome, and can result in large scale CNVs where entire chromosomes have abnormal copy number (Drews et al. 2022). Other mechanisms that may lead to CNVs are replication stress and mitotic errors, as suggested by studies (Sansregret and Swanton 2017; Levine and Holland 2018).

Structural Variation in Tumorigenesis

Structural variants differ from CNVs in size, with SVs generally less than 1Mb and CNV often exceeding this. In addition, SVs can represent differences in DNA quantity as well as changes in location and orientation. SVs identified in short-paired end WGS can be broadly categorised into five main SV types: duplications, deletions, inversion, insertion and translocations. A duplication is an increase in copy number (Figure 2 a) and a deletion is a decrease in copy number (Figure 2 b) of a linear region of the genome. An insertion is the addition of a new region of DNA (Figure 2 c). Inversions and translocations do not alter copy number but instead, an inversion changes the orientation of a region (Figure 2 d), and a translocation changes the location of a region (Figure 2 e).

The prediction and categorisation of SVs relies upon tools that utilise split reads, discordant read pairs and alterations in sequencing depth (Chen et al. 2016; Cameron, Baber, Shale, Papenfuss, et al. 2019). However, the prediction of both SVs and CNVs is an area of active research and no single algorithm is optimally accurate according to published performance comparisons (Cameron, Di Stefano, and Papenfuss 2019). In general researchers generate consensus calls across more than one algorithm to increase accuracy, as used in the recent ICGC PCAWG projects (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisiation/Ins et al. 2020;

Gabrielaite et al. 2021; Moreno-Cabrera et al. 2020).

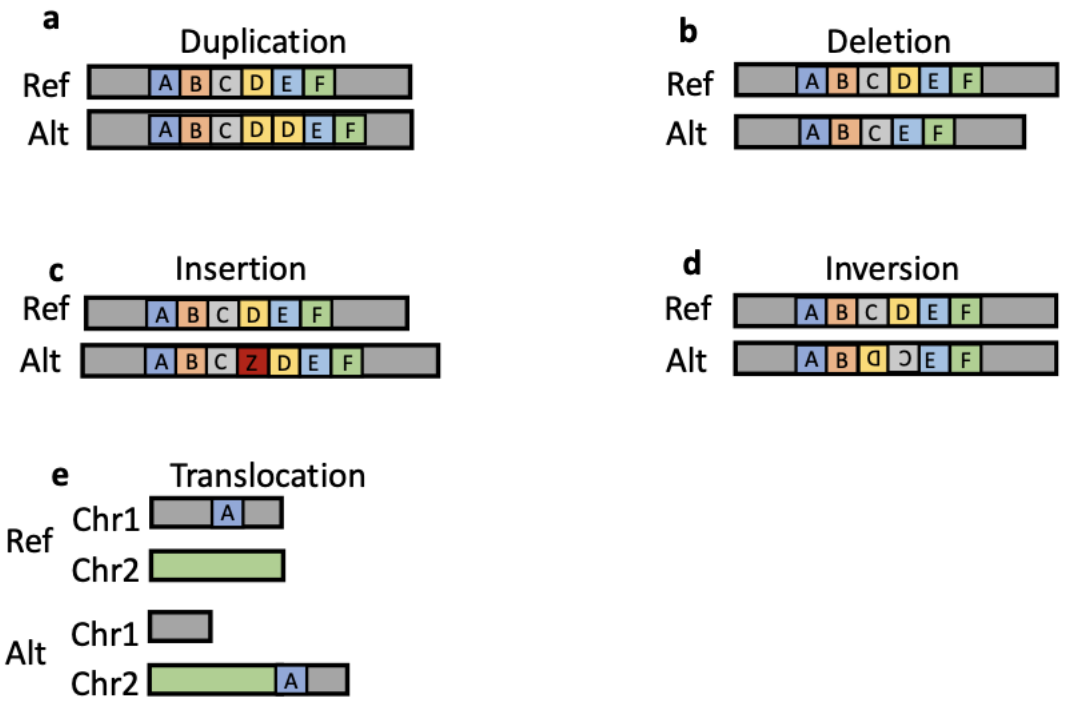


Figure 2 Types of structural variation

The effects of different structural variants on chromosome structure. The panels show the increase or decrease of the amount of DNA caused by duplication and deletion respectively; the addition of DNA from an insertion and the change in orientation or location of DNA from an inversion or translocation respectively. Adapted from Hamdan and Ewing 2022 to show translocations.

The impact of the SVs can depend on the genomic context it occurs in; a deletion in an essential gene may lead to a phenotype, whereas a deletion of the same size in a different location may have no impact on the cell. In fact, a SV was the first genetic alteration to be associated with human cancer with the discovery of the Philadelphia chromosome, generated by a reciprocal translocation in patients with chronic myeloid leukemia (Sabath 2013; Nowell and Hungerford 1960). This translocation creates a fusion gene with increased activity leading to an increase in proliferation (Sabath 2013).

As previously mentioned, SVs have already been shown to be important in HGSOC by impacting the BRAC1 and BRAC2 genes (Ewing et al. 2020). In this thesis, this work will be extended to examine the structural variation affecting all protein coding genes.

The availability of abundant whole genome sequencing data has led to the identification of various types of cSVs.

Figure 3 depicts the different types of cSVs previously reported and their impact on chromosome structure and DNA copy number. The eight known types of cSVs are chromothripsis, extrachromosomal circular DNA (ecDNA), chromoplexy, breakage fusion bridges, pyrgo, rigma, tyfonas, and seismic amplification. In this work, we will delve into

these cSVs in more detai

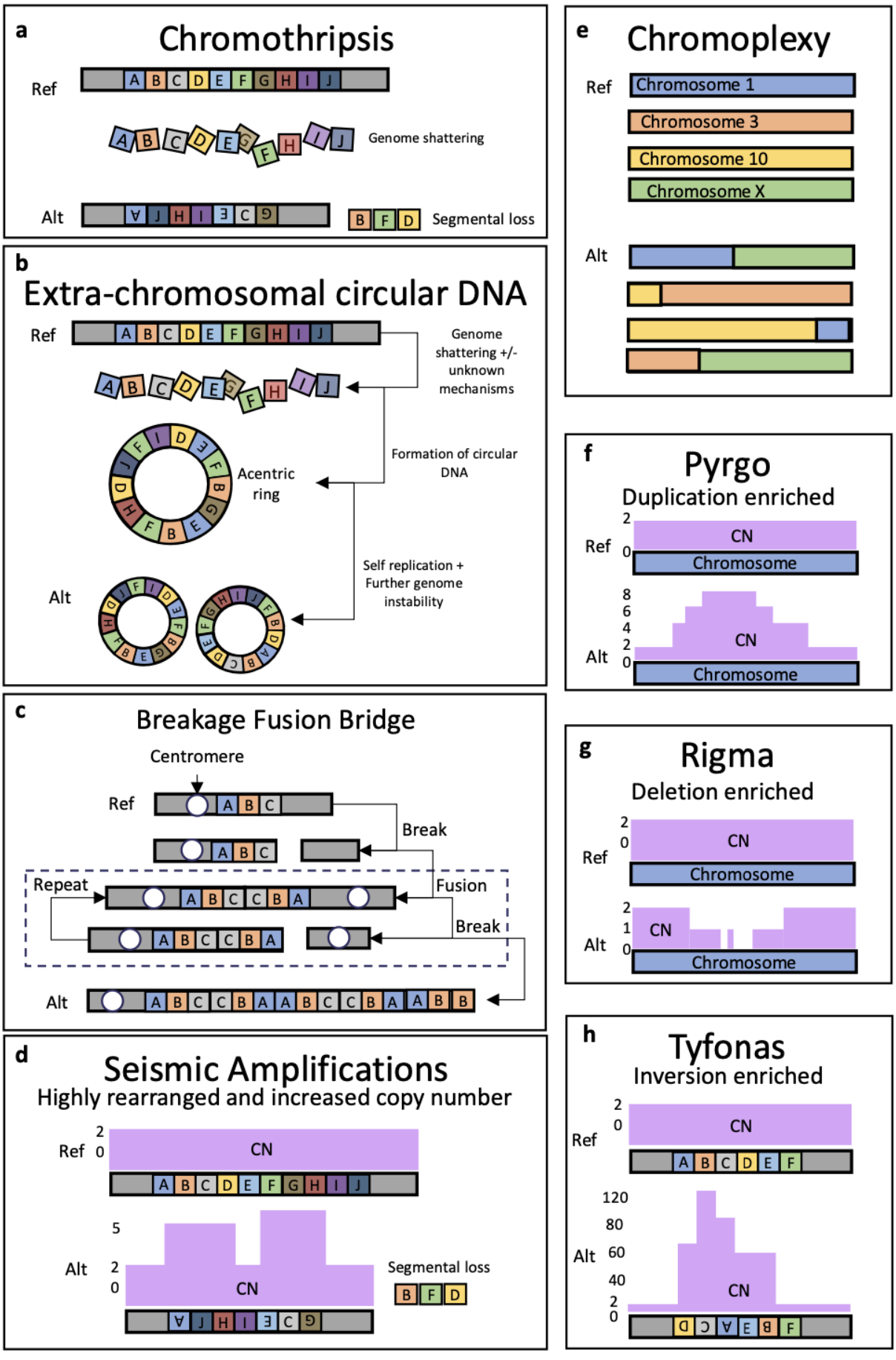


Figure 3 Types of Complex Structural Variants

Previously reported complex structural variants are depicted in successive panels from top left as follows. **a** The shattering and reformation of a chromosome region due to chromothripsis. **b** The rearrangements leading to the formation of the circular structure of ecDNA. **c** The repeated cycles of chromosome breakage and fusion underlying the breakage fusion bridge. The rearranged high copy number regions due to tyfonas (**h**) and seismic amplifications (**d**). **e** The interlinked reciprocal translocations due to chromoplexy. The build-up of duplications or deletions due to prygo (**f**) and rigma (**g**) respectively. Adapted and extended from Hamdan and Ewing 2022 to show chromoplexy, pyrgo, rigma, tyfonas and, seismic amplifications.

Advances In the Identification of Complex Structural Variants

Recently there has been rapid progress in the study of cSVs over the past decade as shown in the timeline in Figure 4 (Shorokhova, Nikolsky, and Grinchuk 2021; Robert and Crasta 2022). Nonetheless, ecDNA was identified even before the first draft of the human reference genome was published. This was done through the use of microscopy to observe metaphase chromosomes, which revealed the presence of small circles of extrachromosomal DNA within cells (Cox, Yuncken, and Spriggs 1965; Lander et al. 2001). These were initially called double minutes due to sometimes being found in pairs (Cox, Yuncken, and Spriggs 1965), though more recent research has found that only 30% of ecDNA occur in pairs (Turner et al. 2017).

Detection of ecDNA

The development of algorithms have played an important role in ecDNA research by standardizing techniques to identify ecDNA. Initially, the algorithm ECdetect was developed to identify ecDNA from DAPI stained metaphase cells (Turner et al. 2017). Using this algorithm, ecDNA was able to be identified in half of all cancer cells and it was observed that ecDNA are almost never found in healthy human cells. (Turner et al. 2017). Built by the same team as ECdetect, the AmpliconArchitect algorithm allows for the identification of ecDNA from combinations of CNV and SV calls derived from WGS data allowing for the identification of ecDNA in tumour samples (Deshpande et al. 2019).

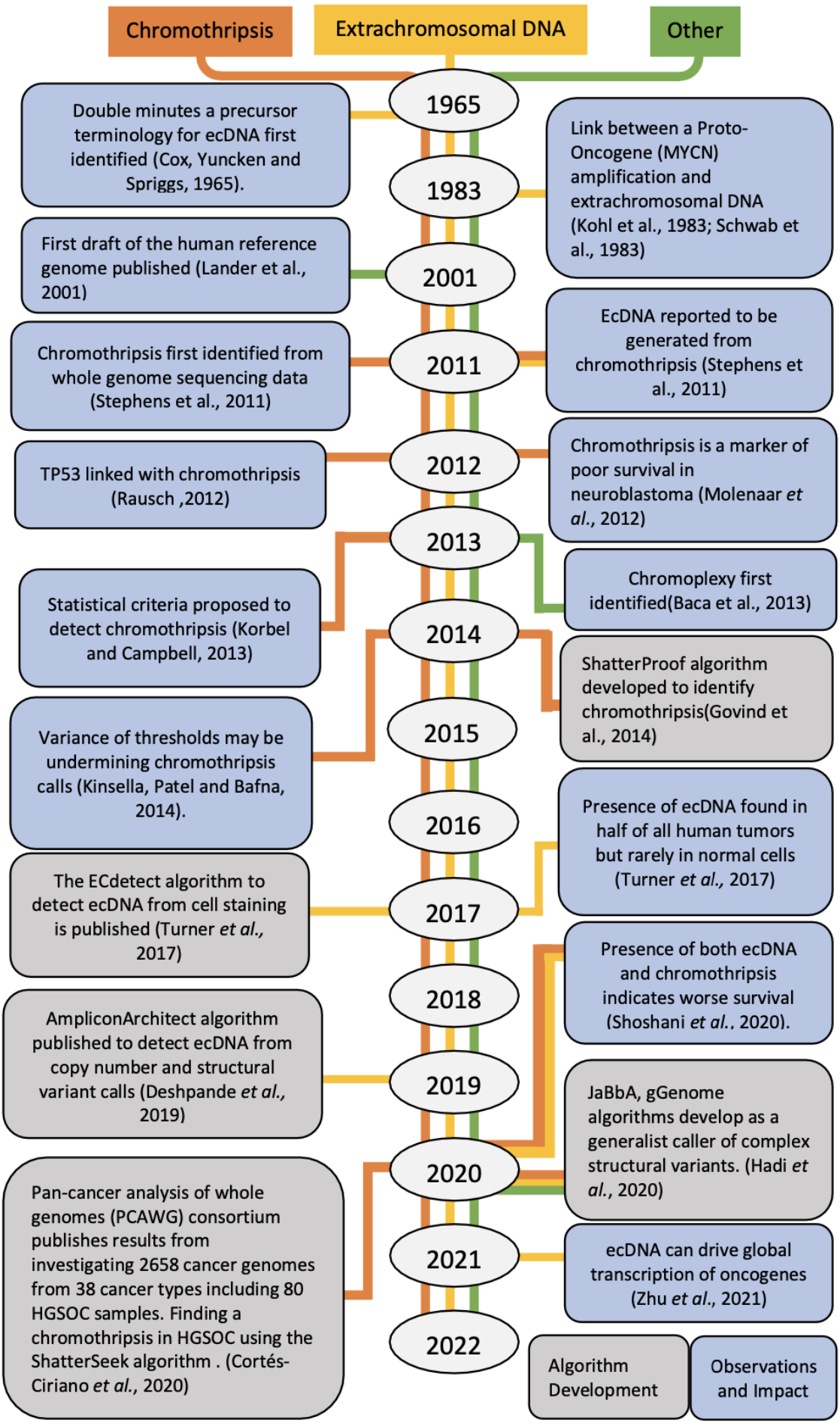

Chromothripsis
Extrachromosomal DNA
Other
1965
Double minutes a precursor terminology for ecDNA first identified (Cox, Yuncken and Spriggs, 1965).
1983
Link between a Proto-Oncogene (MYCN) amplification and extrachromosomal DNA (Kohl et al., 1983; Schwab et al., 1983)
2001
First draft of the human reference genome published (Lander et al., 2001)
2011
Chromothripsis first identified from whole genome sequencing data (Stephens et al., 2011)
EcDNA reported to be generated from chromothripsis (Stephens et al., 2011)
2012
TP53 linked with chromothripsis (Rausch ,2012)
Chromothripsis is a marker of poor survival in neuroblastoma (Molenaar et al., 2012)
2013
Statistical criteria proposed to detect chromothripsis (Korbel and Campbell, 2013)
Chromoplexy first identified(Baca et al., 2013)
2014
ShatterProof algorithm developed to identify chromothripsis(Govind et al., 2014)
2015
Variance of thresholds may be undermining chromothripsis calls (Kinsella, Patel and Bafna, 2014).
2016
Presence of ecDNA found in half of all human tumors but rarely in normal cells (Turner et al., 2017)
2017
The ECdetect algorithm to detect ecDNA from cell staining is published (Turner et al., 2017)
2018
Presence of both ecDNA and chromothripsis indicates worse survival (Shoshani et al., 2020).
2019
AmpliconArchitect algorithm published to detect ecDNA from copy number and structural variant calls (Deshpande et al., 2019)
2020
JaBbA, gGenome algorithms develop as a generalist caller of complex structural variants. (Hadi et al., 2020)
Pan-cancer analysis of whole genomes (PCAWG) consortium publishes results from investigating 2658 cancer genomes from 38 cancer types including 80 HGSOC samples. Finding a chromothripsis in HGSOC using the ShatterSeek algorithm . (Cortés-Ciriano et al., 2020)
2021
ecDNA can drive global transcription of oncogenes (Zhu et al., 2021)
2022
Algorithm Development
Observations and Impact

Figure 4 Timeline of milestones in complex structural variant biology

The timeline highlights key developments in algorithm development and novel insights into complex structural variant biology. Gray boxes show algorithmic development and blue box show observations and impact.

Detection of Chromothripsis

The generation of ecDNA has been linked to another cSV, chromothripsis (Stephens et al. 2011; X.-K. Zhao et al. 2021a). Chromothripsis involves the shattering and random reforming of one or a few chromosomes (Stephens et al. 2011). First named and discovered in patients with chronic lymphocytic leukaemia, with the distinct patterns and density of genomic rearrangements observed thought unlikely to have developed sequentially (Stephens et al. 2011). This pattern was instead proposed to have occurred within a single cell cycle, indicating that chromothripsis did not fit the classical mutational theory of oncogenesis, where there is a gradual build-up of mutations over many cell cycles (Stephens et al. 2011). This pattern of rearrangement was localised within the genome, with alternating copy number caused by an approximately even number of deletions and duplications, as well as inversions and translocations also clustering in these regions (Stephens et al. 2011).

To support the statement that the pattern observed was the result of shattering and reforming of the chromosome in a single cell cycle, Monte Carlo simulations were used (Stephens et al. 2011). The simulations of the progressive rearrangement model were performed 1000 times randomly sampling from 239 structural variants that they had identified in a human sample and measuring the predicted copy number states (Stephens et al. 2011). From these simulations, it was determined that the observed pattern was unlikely to occur from progressive accumulation of rearrangements and thus the structural variants were likely to occur simultaneously (Stephens et al. 2011).

Chromothripsis has been reported in the literature using different numbers (such as 20, 10 or 6) of structural variants as thresholds to indicate chromothripsis (Molenaar et al. 2012; Northcott et al. 2012; Rausch et al. 2012; Chiang et al. 2012). It has been noted that this variation in thresholds may affect the ability to distinguish simultaneous chromothripsis SV generation from sequential SV accumulation (Kinsella, Patel, and

Bafna 2014). Kinsella et al used statistical simulations to show that reducing thresholds increased the chance that complex rearrangements are identified by random change (Kinsella, Patel, and Bafna 2014). This could result from sequential rearrangement and therefore increasing the chances of misidentifying regions as affected by chromothripsis (Kinsella, Patel, and Bafna 2014).

To address such difficulties in identifying chromothripsis Korbel and Campbell proposed statistical criteria for identification of chromothripsis based on simulated data (Korbel and Campbell 2013). The criteria proposed were as follows: clustering of breakpoints, randomness of DNA fragment joins and randomness of DNA fragment order (Korbel and Campbell 2013).

Clustering of breakpoints was proposed to be tested by comparing the distance between the genomic positions of adjacent breakpoints of breakpoints to a null random distribution of breakpoints (Korbel and Campbell 2013).

The randomness of DNA joins by counting the frequency of the four different types of joins (deletion, tandem duplication, head-to-head-inverted, and tail-to-tail-inverted) and comparing them to an expected frequency of 25**%** with departure from this frequency being evidence against chromothripsis. This is testing that assumption that fragments of the shattered chromosomes are joined together randomly (Korbel and Campbell 2013).

The randomness of DNA fragment order can be tested by indexing breakpoints by genomic position then by comparing the index of breakpoints linked by SV pairs should be random and drawn without replacement. Korbel and Campbell suggest that the randomness can be tested in multiple ways including Monte Carlo simulations. In practice Korbel and Campbell found that the DNA fragment order was not truly random but more random than would be expected by a progressive model of rearrangement which would be the alternative to shattering and reforming. The randomness of DNA fragment order tests the assumption that the shattered fragments of DNA are recombined randomly.

The first algorithm published to call chromothripsis was ShatterProof (Govind et al.

2014), but the criteria used by ShatterProof to identify chromothripsis were not based on those proposed by Korbel and Campbell. Instead copy number variation, translocation, insertion and loss of homology calls were used to produce a weighted score of how likely a highly mutated region is to have involved chromothripsis (Govind et al. 2014). The weighting of copy-number aberrations and translocations as joint highest made these the two the most important criteria for calling chromothripsis (Govind et al. 2014). Despite being an important step toward standardizing the identification of chromothripsis, ShatterProof was not used in the ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) study which investigated 2,658 tumour samples for the presence of genomic features including chromothripsis (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020).

An alternative algorithm, ShatterSeek, which was based on the statistical criteria proposed by Korbel and Campbell was used for the PCAWG study (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). As part of the PCAWG study the results of the ShatterSeek and ShatterProof algorithms were compared in 109 prostate adenocarcinoma samples which had been previously tested for chromothripsis by the authors of ShatterProof, who identified chromothripsis in 23 (21%) samples with ShatterSeek (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020; Govind et al. 2014). However, this rose to 49 (45%) samples when the ShatterProof analysis was redone with the more reliable PCAWG consensus copy number and SV calls (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). ShatterSeek, identified 60 (55%) samples as having chromothripsis. In addition, 4 of the samples identified by ShatterProof were not identified as having chromothripsis by ShatterSeek (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020), as the regions concerned had less than 6 SVs involved which is below the minimum threshold ShatterSeek sets for identification of chromothripsis (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020).

ShatterSeek was also compared to another algorithm, ChromAL, in 76 pancreatic tumours and both detected the same 41 (54%) samples as affected by chromothripsis (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). ChromAL was also based on the same statistical criteria for calling chromothripsis proposed by Korbel and Campbell (Notta,Chan,2016). ChromAL and ShatterSeek have very similar criteria, but

ShatterSeek provides both high and low confidence calls of chromothripsis and was the preferred algorithm for PCAWG analyses (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020; Notta et al. 2016).

Detection of Chromoplexy

As WGS data has accumulated, more detailed studies of tumour mutational spectra have revealed cSVs beyond chromothripsis, such as chromoplexy (M. M. Shen 2013).

Chromothripsis and chromoplexy are two genomic rearrangement processes that involve random breakage and fusion of genomic segments, resulting in changes in copy number states. However, there are some key differences between the two processes (Baca et al. 2013). Chromothripsis typically generates hundreds of clustered structural variant (SV) breakpoints and is usually restricted to one or two chromosomes. In contrast, chromoplexy typically generates tens of SVs, which are 'chained' together as a complex series of rearrangements, rather than clustered across specific regions, and can involve multiple chromosomes (Baca et al. 2013).

Chromoplexy was originally identified in prostate cancer and was proposed to be a driver event generating oncogenic gene fusions and tumour suppressor losses (Baca et al. 2013).The identification of chromoplexy relies on detecting chained together SVs across multiple chromosomes and as such it cannot be identified when regions of the genome are considered individually (Baca et al. 2013). This means that chromoplexy detection is well suited to algorithms rooted in graph theory which treats information as a set of nodes and arcs this then allows for circular paths to be identified. This is the approach used by Chainfinder and subsequently gGenomes (Baca et al. 2013; Hadi et al. 2020).

Detection of Rigma, Pyrgo and Tyfonas

A key statement in the PCAWG paper on chromothripsis was that as the accuracy of CNV and SV calls has improved so too has the ability to call cSVs (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). An important development in the ability to call CNVs and SVs

came from the combined use of GRIDSS, PURPLE, LINX which was designed to be an integrated and comprehensive structural variant and copy number analysis tool kit and also provides purity and ploidy estimates (Cameron, Baber, Shale, Papenfuss, et al. 2019). By using changes in copy number to support its SV calls it was able to rescue SV calls that without that additional support would not have been identified (Cameron, Baber, Shale, Papenfuss, et al. 2019). Although still in preprint at time of writing GRIDSS, PURPLE, LINX has been utilised as the input to a generalist cSV caller of JaBbA and gGenome (Cameron, Baber, Shale, Papenfuss, et al. 2019; Hadi et al. 2020).

The Junction Balance Analysis (JaBbA) algorithm combines coverages and SV breakpoints (junctions) calls from GRIDSS, PURPLE, LINX to build a genome graph. This graph is then used by gGenome to classify cSV.

The combination of JaBbA and gGenome also allowed for the identification of many types of previously identified cSV types including chromothripsis, chromoplexy and ecDNA (Hadi et al. 2020). Building on the development in the GRIDSS, PURPLE, LINX pipeline and utilising a novel genome graph algorithm approach, the authors of gGenome algorithms reported novel complex structural variants named rigma, pyrgo and tyfonas.

Rigma was identified as an event generating clusters of deletions with reported associations with late replicating regions, fragile sites and long genes when compared to simple deletions (Hadi et al. 2020). Similarly, pyrgo was identified as generating clusters of duplications, with a reported association with early replicating regions and super-enhancers when compared to simple duplications (Hadi et al. 2020). These differences in their genomic distributions were proposed to be the main evidence that rigma and pyrgo arise from different mutational processes and or selection pressures than simple deletions or duplications. Tyfonas generates amplified copy number inversions and was proposed to differ from other cSVs that involve amplified copy number and inversions such as breakage fusion bridges due to their enrichment in particular tumour types (Hadi et al, 2020).

Detection of Breakage Fusion Bridges

The mechanism for the generation of breakage fusion bridges (BFBs) proposed involves the loss of a telomere (McClintock 1941; Umbreit et al. 2019). During mitosis the two copies of the chromatids lacking the telomere fuse to create an aberrant chromosome with two centromeres (McClintock 1941; Umbreit et al. 2019). This aberrant chromosome still lacks telomeres allowing further BFB cycles to continue and produce further genomic instability (McClintock 1941; Zakov, Kinsella, and Bafna 2013; Umbreit et al. 2019). The identification of BFBs relies on the identification of the increase in copy number and inversions caused by the repeated fusion, breakages and duplications of chromosomes over multiple cell cycles (Hadi et al. 2020; Baca et al. 2013).

Detection of Seismic Amplification

The massive rearrangements involved in chromothripsis have been proposed to be the source of ecDNA and also other cSV types (Rosswog et al. 2021; X.-K. Zhao et al. 2021a). A recent study of 79 WGS neuroblastoma samples found amplification (copy number >= 5 in diploid and >= 9 in polyploid genomes) in 57 samples (72%). Rosswog et al also reported that 19 (24%) of these amplified samples also contained a new type of cSV in the same amplified regions (Rosswog et al. 2021). This pattern of cSV was characterised by at least 14 internal rearrangements and elevated copy number, and was termed seismic amplification (Rosswog et al. 2021).

Seismic amplification differed from chromothripsis in that it lacked the characteristic oscillation of copy number that is characteristic of chromothripsis, and differed from other amplifications due to the complexity of the rearrangements involved (Rosswog et al. 2021). Using fluorescent *in situ* hybridisation, these regions of seismic amplification were observed to exist extra-chromosomally and integrated into chromosomes (Rosswog et al. 2021). It was suggested that chromothripsis may be involved in the generation of seismic amplifications that possess dense clustering of breakpoints and uniform distribution of SV types which are both hallmarks of chromothripsis (Rosswog et al. 2021; Stephens et al. 2011; Korbel and Campbell 2013).

Additionally, it was observed that chromothripsis occurred in 89% of samples (244 of 274 samples from multiple tumour types) showing evidence for seismic amplification. (Rosswog et al. 2021) However, simulation studies found that chromothripsis followed by cyclic amplification most closely matched the copy number and SV patterns observed in seismic amplification identified in the PCAWG cohort (Rosswog et al. 2021). These findings suggest that while chromothripsis and ecDNA may play a significant role in the development of seismic amplification, it may not be the only mechanism involved.

Functional and Clinical Impacts of Complex Structural Variants

Many cSV events have been described as genomic chaos or crisis that generate variation which can be exploited by selection during tumorigenesis (G. Liu et al. 2014). In the same way that the genomic context of a SV will affect its impact so will the genomic context of a cSV. Different types of cSV alter the regions they impact in different ways, for example, the amplification of copy number in a region by ecDNA is likely to have very different consequences from the oscillation of copy number across a region affected by chromothripsis. Although many cSV have only recently been reported and therefore the literature on their impact is limited, there is an emerging literature on their functional and clinical impacts.

Impact of ecDNA on Patients

The potential importance of ecDNA in tumour biology was established when the MYC oncogene was linked to ecDNAs in two separate studies in 1983, suggesting ecDNA oncogene amplification as a tumour driver mechanism (Kohl et al. 1983; Schwab et al. 1983). Subsequently, ecDNA have been found to frequently contain oncogenes in a number of tumour types, and the presence of ecDNA has been associated with worse patient outcome (Kim et al. 2020; Turner et al. 2017; Verhaak, Bafna, and Mischel 2019). Using live cell imaging of cells containing ecDNA, the uneven segregation of ecDNA has been observed during mitosis, meaning that the number of ecDNA in the daughter cells can vary (Yi et al. 2022). This variation can lead to rapid tumor evolution when proliferative ecDNA containing clonal lineages have a selective advantage (Yi et al. 2022). The uneven inheritance of ecDNA allows for rapid changes in copy number to quickly find the optimal level for ecDNA with beneficial or detrimental phenotypes.

In addition, ecDNA has been subject to various speculative mechanisms to explain its functional impact. One such mechanism suggests that ecDNA tends to spatially cluster within cells, which can increase the expression of genes contained on the ecDNA. This clustering may occur because of increased spatial proximity, leading to more efficient use of transcription machinery (Hung et al. 2021; Purshouse et al. 2022).

Recently, ecDNA were observed to act as mobile regulatory elements driving global transcription of oncogenes. Meaning that ecDNA might impact the expression of chromosomal genes as well as those resident on the ecDNA by clustering within the nucleus (Zhu et al. 2021). However, experiments using high resolution microscopy have suggested that the simple increase in copy number facilitated by ecDNAs is the main factor explaining the increase in oncogene expression associated with ecDNA (Purshouse et al. 2022).

Impact of Chromothripsis on Patients

Chromothripsis has been associated with poor patient outcome in several cancer types including medulloblastoma, neuroblastoma, colorectal cancer and myeloid leukaemia (Voronina et al. 2020; Waszak et al. 2018; Molenaar et al. 2012; Fontana et al. 2018; Chiara et al. 2018). The co-occurrence of ecDNA and chromothripsis was also reported to be associated with poor patient outcomes (Shoshani et al. 2020). However, the association with worse survival has been inconsistent across tumour types, and was not found in colorectal cancer or uveal melanomas (Skuja et al. 2017; van Engen-van Grunsven et al. 2015). In contrast, chromothripsis has also been reported to have a positive influence on progression free survival in metastatic colorectal cancer (Skuja et al. 2017). As mentioned above, many cSV studies have differed in the criteria used to identify chromothripsis, and this problem also affects these studies of clinical impact (Voronina et al. 2020; Waszak et al. 2018; Molenaar et al. 2012; Fontana et al. 2018).

Impact of Chromoplexy on Patients

Chromoplexy affecting the activity of super enhancers and oncogenes has been reported to have a negative impact on progression free survival in multiple myeloma (Ashby et al. 2018). In addition, chromoplexy has been reported to enable ERG fusions or dysregulation and PTEN loss, the combination of which has been shown to trigger the formation of aggressive prostate tumours, implying that chromoplexy has a negative impact on survival (Berger et al. 2011; Carver et al. 2009; King et al. 2009).

Impact of Breakage Fusion Bridges on Patients

A combination of fluorescence *in situ* hybridization and WGS studies have shown that BFBs can increase the numbers of SVs found on chromosomes and can act as a source of variation upon which selective pressure can act and clonal evolution can take place (Gisselsson et al. 2000). In mice with mutations in TP53 and nonhomologous end joining DNA repair proteins, lymphomas have been reported with BFB events involving amplification of the MYC oncogene (Umbreit et al. 2019; Difilippantonio et al. 2002).

Aims

Simple SVs and SNVs have been shown to impact tumour gene expression and patient outcomes in many tumour types, yet the recurrent patterns of SV in highly rearranged cancer types such as HGSOC are relatively understudied (Ewing et al. 2020; Cosenza, Rodriguez-Martin, and Korbel 2022). Multiple types of cSV have also been identified as, discussed above, but their impact and importance in HGSOC is unknown. In addition, many cSV studies have focused on a single type of cSV so their patterns of co-occurrence have been understudied. Finally, the effects of different cSV types on the function of the genes they impact and on patient survival are unknown in HGSOC. It is these questions that this work will further investigate with the following specific aims.

1. Determine the diversity, recurrence, frequency and interaction of structural variant and complex structural variants across the combined cohort of HGSOC.

2. Investigate the co-occurrence and mutual exclusivity of complex structural variation in HGSOC, with different cSV types and other relevant genomic features.

3. Assess the functional impact of structural variants and complex structural variants on the genes they cover.

4. Determine the impact of structural variants and complex structural variants on patient survival time and time to relapse and assess the value of structural variants and complex structural variants as biomarkers of patient outcomes.

Chapter 2: Methods

Calling Structural Variants and Copy Number Variants

Somatic structural variants (SVs) and copy number variants (CNVs) were identified by other members of the Semple and Ewing labs using a consensus approach with specialized callers. For consensus SVs, the viola-sv tool was employed, applying a proximity threshold of 100 bp to determine the intersection of SV calls from Manta and GRIDSS (Cameron, Baber, Shale, Papenfuss, et al. 2019; Chen et al. 2016; Sugita et al. 2022). Similarly, consensus CNVs were generated by considering a minimum 50% overlap between segments identified by CNVkit, CLImAT, and PURPLE (Talevich et al. 2016; Cameron, Baber, Shale, Papenfuss, et al. 2019; Yu et al. 2014).

Identifying Complex Structural Variants

Chromothripsis

To identify chromothripsis within the combined cohort the R package ShatterSeek was utilised (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020). Based on the statistical criteria proposed by Korbel and Campbell, ShatterSeek provides criteria for identifying chromothripsis (Korbel and Campbell 2013). From manta SV calls and CNVkit copy number calls ShatterSeek was used to identify chromothripsis from the number of interleaved SVs, and copy number oscillations, as well as chromosomal enrichment of breakpoints and random fragment joining to predict chromothripsis (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020). An example of an interleaved SV is shown in Figure 5 and is defined as an SV that has one breakpoint between the breakpoints of another SV. Notable this does not addresses the problem of SV phasing.

Phasing of SV is where it is attempted to determine if two rearrangements events occur on the same copy of a chromosome in a diploid genome it is possible if two events occur on the same or different copies of a chromosome. This problem is made increasingly difficult by whole genome doubling where the number of copies of each chromosome double and therefore the number of different places SV could be increase. With short read sequencing where the reads are ~150 bp there is no information about the phasing of breakpoints more than 150 bp from

each other. Long read sequencing increases the distance between breakpoint for which phasing information is available. The phasing of SV is an ongoing challenge in the field of structural variant research and is beyond the scope of this thesis.

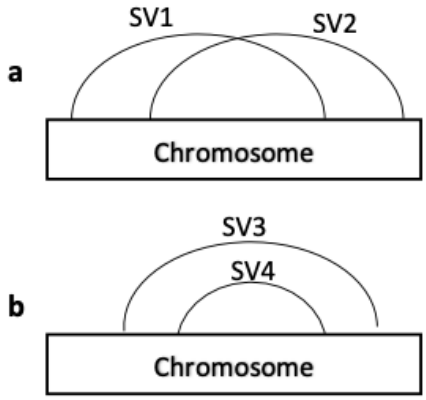
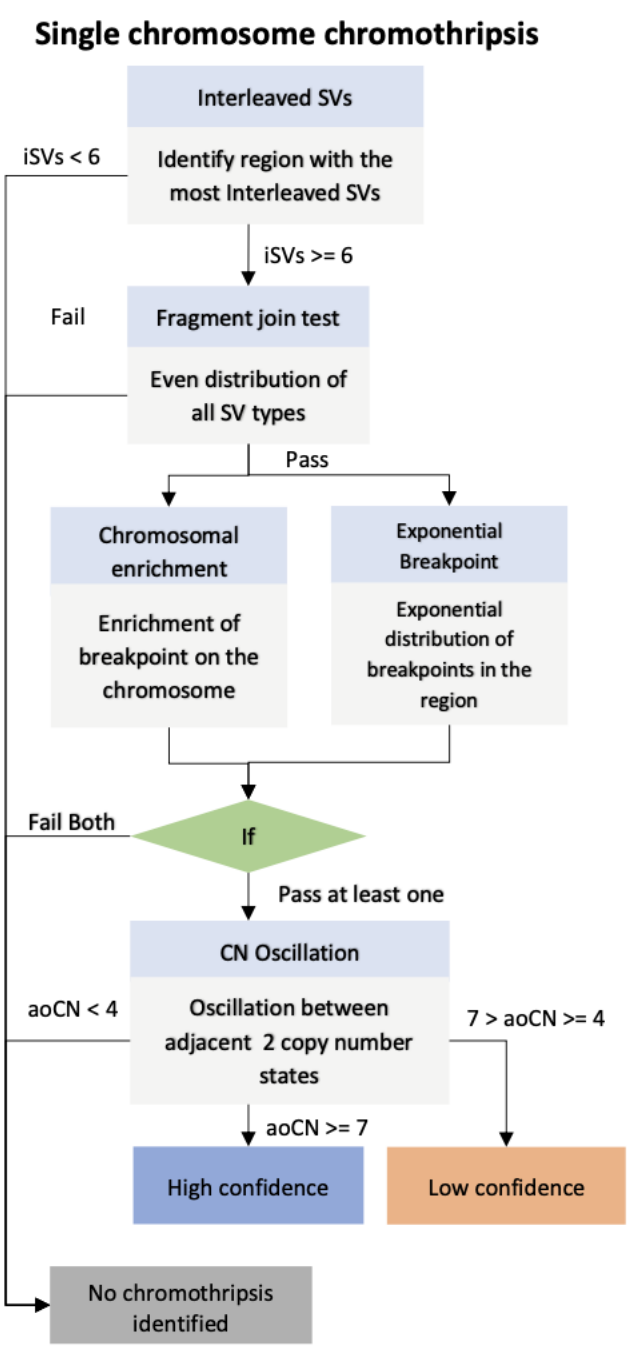
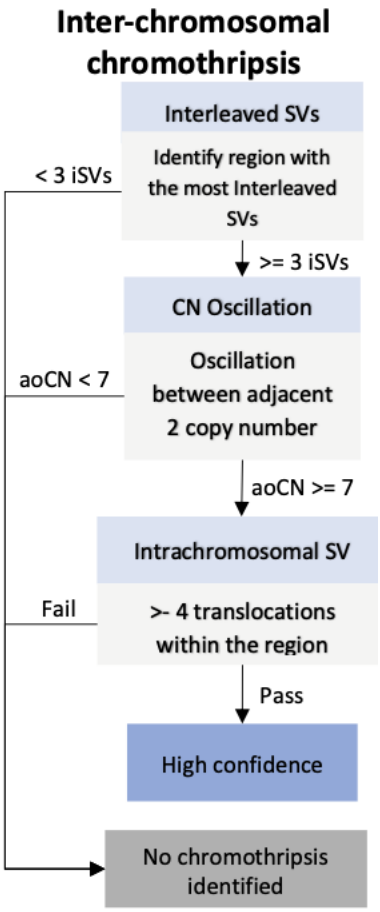


Figure 5 Interleaved structural variant

Each arch above represents a structural variant (SV). **a** shows a pair of interleaved SVs with one end of SV1 between each end of SV2. **b** shows a pair of SVs as both ends of SV4 are between the ends of SV3 these SVs are nested and are not interleaved.


Single chromosome chromothripsis
Interleaved SVs
Identify region with the most Interleaved SVs
iSVs < 6
iSVs >= 6
Fail
Fragment join test
Even distribution of all SV types
Pass
Chromosomal enrichment
Enrichment of breakpoint on the chromosome
Exponential Breakpoint
Exponential distribution of breakpoints in the region
Fail Both
If
Pass at least one
CN Oscillation
aoCN < 4
Oscillation between adjacent 2 copy number states
7 > aoCN >= 4
aoCN >= 7
High confidence
Low confidence
No chromothripsis identified


Inter-chromosomal chromothripsis
Interleaved SVs
Identify region with the most Interleaved SVs
< 3 iSVs
>= 3 iSVs
CN Oscillation
Oscillation between adjacent 2 copy number
aoCN < 7
aoCN >= 7
Intrachromosomal SV
Fail
>- 4 translocations within the region
Pass
High confidence
No chromothripsis identified

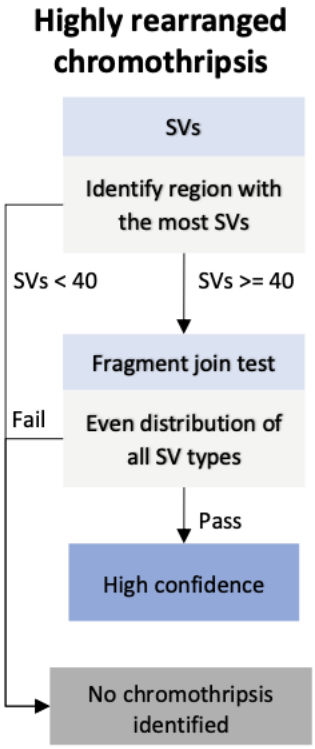

Highly rearranged chromothripsis
SVs
Identify region with the most SVs
SVs < 40
SVs >= 40
Fragment join test
Fail
Even distribution of all SV types
Pass
High confidence
No chromothripsis identified

Figure 6 Flow diagram for identification of chromothripsis by ShatterSeek

A flow diagram for the identification of chromothripsis from structural variant and copy number calls using the ShatterSeek criteria (Cortés-Ciriano et al. 2020). There are three routes to identify chromothripsis: chromothripsis on one chromosome, chromothripsis identified on multiple chromosomes, and highly structurally rearranged chromothripsis. Acronyms used in the figure are; Structural variant (SV), interleaved SV (iSV), and adjacent oscillation of copy number (aoCN).

ShatterSeek attempt to identify chromothripsis found in multiple context as shown by the three routes in Figure 6 to reach a call of high confidence chromothripsis. The first route is a region with at least 40 SVs with roughly equal proportions of each SV type, this is to identify chromothripsis in highly rearranged regions missed by other routes (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020). The second route to high confidence classification aims to identify chromothripsis spread across multiple chromosomes by considering inter-chromosomal SVs and as such includes intrachromosomal SV as an key measure (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020). The third route to confidence classification aims to identify chromothripsis located on one chromosome requiring more interleaved SVs than the second route (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020). The low confidence classification differs from the third route based on the number of adjacent segment oscillating copy number states (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020).

Chromothripsis was also called by the gGenome (Hadi et al. 2020). Although based on the criteria proposed by Korbel and Campbell, gGenome attempts to identify chromothripsis using different criteria than ShatterSeek. The criteria used by gGenome (Figure 7) is most similar to the criteria used in the single chromosome chromothripsis route in ShatterSeek.

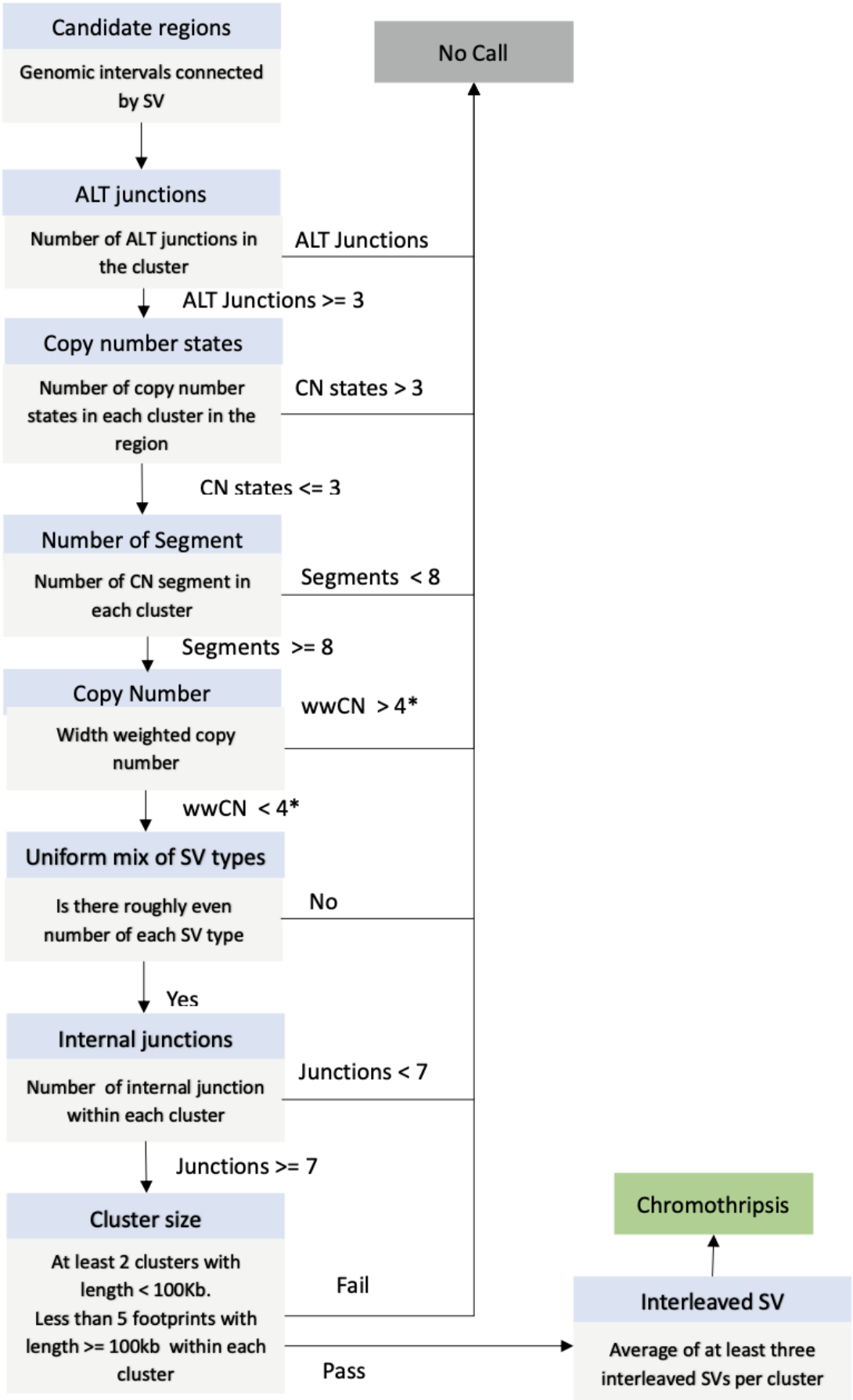

Candidate regions
Genomic intervals connected by SV
No Call
ALT junctions
Number of ALT junctions in the cluster
ALT Junctions
ALT Junctions >= 3
Copy number states
Number of copy number states in each cluster in the region
CN states > 3
CN states <= 3
Number of Segment
Number of CN segment in each cluster
Segments < 8
Segments >= 8
Copy Number
Width weighted copy number
wwCN > 4*
wwCN < 4*
Uniform mix of SV types
Is there roughly even number of each SV type
No
Yes
Internal junctions
Number of internal junction within each cluster
Junctions < 7
Junctions >= 7
Cluster size
At least 2 clusters with length < 100Kb.
Less than 5 footprints with length >= 100kb within each cluster
Fail
Pass
Chromothripsis
Interleaved SV
Average of at least three interleaved SVs per cluster

Figure 7 Flow diagram for identification of chromothripsis by gGenome

A flow diagram for the identification of chromothripsis from structural variant and copy number calls by gGenome.* The threshold for width weighted copy is 4 or twice the sample ploidy whichever is greater.

Although both ShatterSeek and gGenomes attempt to implement the criteria set out by Korbel and Campbell, the criteria is implemented differently. For example in the identification of oscillation of copy number, ShatterSeek identifies the maximum number of adjacent oscillations between two copy number states whereas gGenomes restricts the number of copy number states in a region and then requires a minimum number of changes in copy number. Both of these approaches will identify oscillation in copy number and there is currently no standard or consensus on which way is best. However, gGenome also introduces criteria not recommended by Korbel and Campbell such as restricting the size of regions that can be included in a call of chromothripsis. Additionally at the time of writing Jabba and gGenome are currently in preprint so have not completed the peer review process (Hadi et al. 2020; Korbel and Campbell 2013).

In the absence of a gold standard reference data upon which both callers could be compared an informed choice had to be made between them; Ultimately, ShatterSeek was used in this work to identify chromothripsis as it more closely adheres to the Korbel and Campbell criteria, has been peer reviewed and used in landmark work in cSV literature (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020).

Extrachromosomal circular DNA

To identify extrachromosomal circular DNA (ecDNA) the AmpliconArchitect package was used to identify amplicons and ecDNA and other patterns of SVs (Deshpande et al. 2019). AmpliconArchitect uses short-paired end whole genome alignment generated by other members of the Semple lab, and structural variants called using the structural variant caller, PURPLE, generated by other members of the Semple and Ewing labs (Cameron, Baber, Shale, and Papenfuss 2019). AmpliconArchitect identifies amplicons as genomic regions with increased copy number linked by SVs. In addition to be classed as

amplicons at least one segment of the amplicon must have a copy number >5 and the total amplicon length must be >100 kbp.

The results of the AmpliconArchitect package were classified using AmpliconClassifier package into three categories ; Linear amplification, Complex non-cyclic and, Cyclic (Deshpande et al. 2019; Kim et al. 2020). Linear amplification were regions of the genome with at least one segment with a copy number >4 and no SV or SVs linking regions <1Mb apart and not forming a circular structure. Complex non-cyclic regions are similar to Linear amplification except there is at least one SV linking regions that are >1Mb apart in the reference genome. Cyclic regions are divided into two sub groups, breakage-fusion-bridge (BFB) or ecDNA. The BFB classification was called if more than 25% of discordant reads form fold-back orientation. An ecDNA was called when the segments formed a loop linked by SVs, a size greater than 10kb and a copy number greater than 4. Additionally, AmpliconClassifier predicts a structure of cyclic DNA by combining amplified regions linked by SVs with the same predicted copy number.

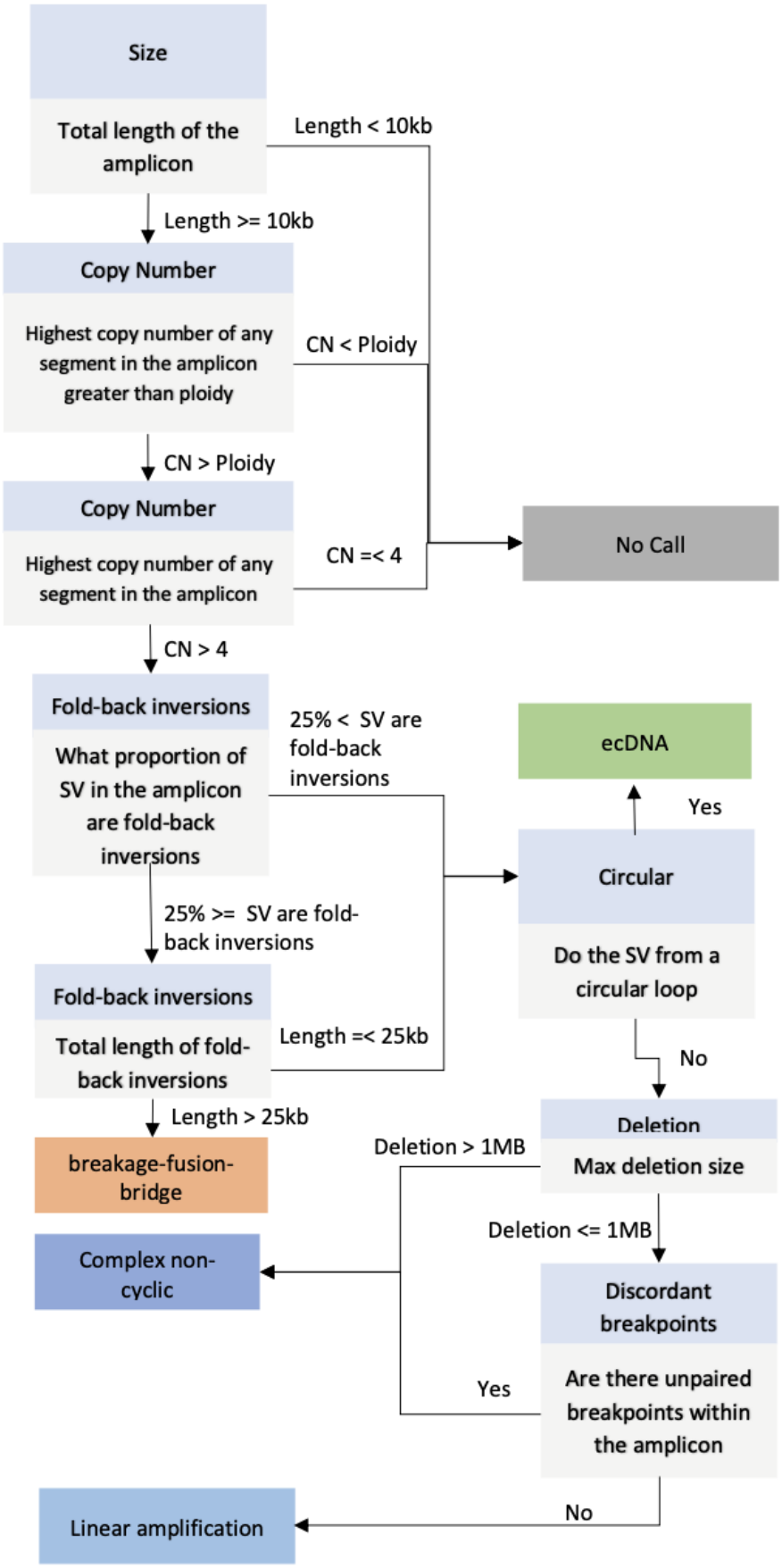


Figure 8 Flow diagram for identification of ecDNA by AmpliconClassifier

The flow diagram demonstrates how AmpliconClassifier classifies amplicons generated by AmpliconArchitect into five categories bases on size, copy number, proportion of fold-back inversion SV, the circularity of the amplicon, deletion size and the presence of discordant reads. The five possible classifications of an AmpliconArchitect amplicon are; Breakage Fusion Bridge, ecDNA, Complex non cyclic, linear amplification and no call.

Rigma and Pyrgo

The criteria for identification of rigma and prygo are nearly identical with the exception that they are a clustering of deletion or duplications respectively (Hadi et al. 2020). Rigma and pygo were identified using the JaBba gGenomes pipeline from PURPLE junction calls and coverage. The criteria for their identification is shown in Figure 9 a and Figure 9 b respectively.

Both Rigma and Pyrgo define their region by looking for overlapping structural variants (deletion in rigma and duplication in pyrgo). They require the regions copy number to be different from the overall ploidy of the sample decreased copy number in rigma and increase in pyrgo (Hadi et al. 2020). Candidate regions with a length less than 10kb or greater than 1Mb are then rejected (Hadi et al. 2020). Finally the region must be enriched for their respective SV type, deletion in rigma and duplication in pyrgo, this enrichment is tested using the fishhook algorithm using a Poisson model to see if the region is enriched for the respective SV type while counting for the presence of other SVs (Imielinski, Guo, and Meyerson 2017; Hadi et al. 2020). The threshold for significance was a FDR <0.5 (Hadi et al. 2020).

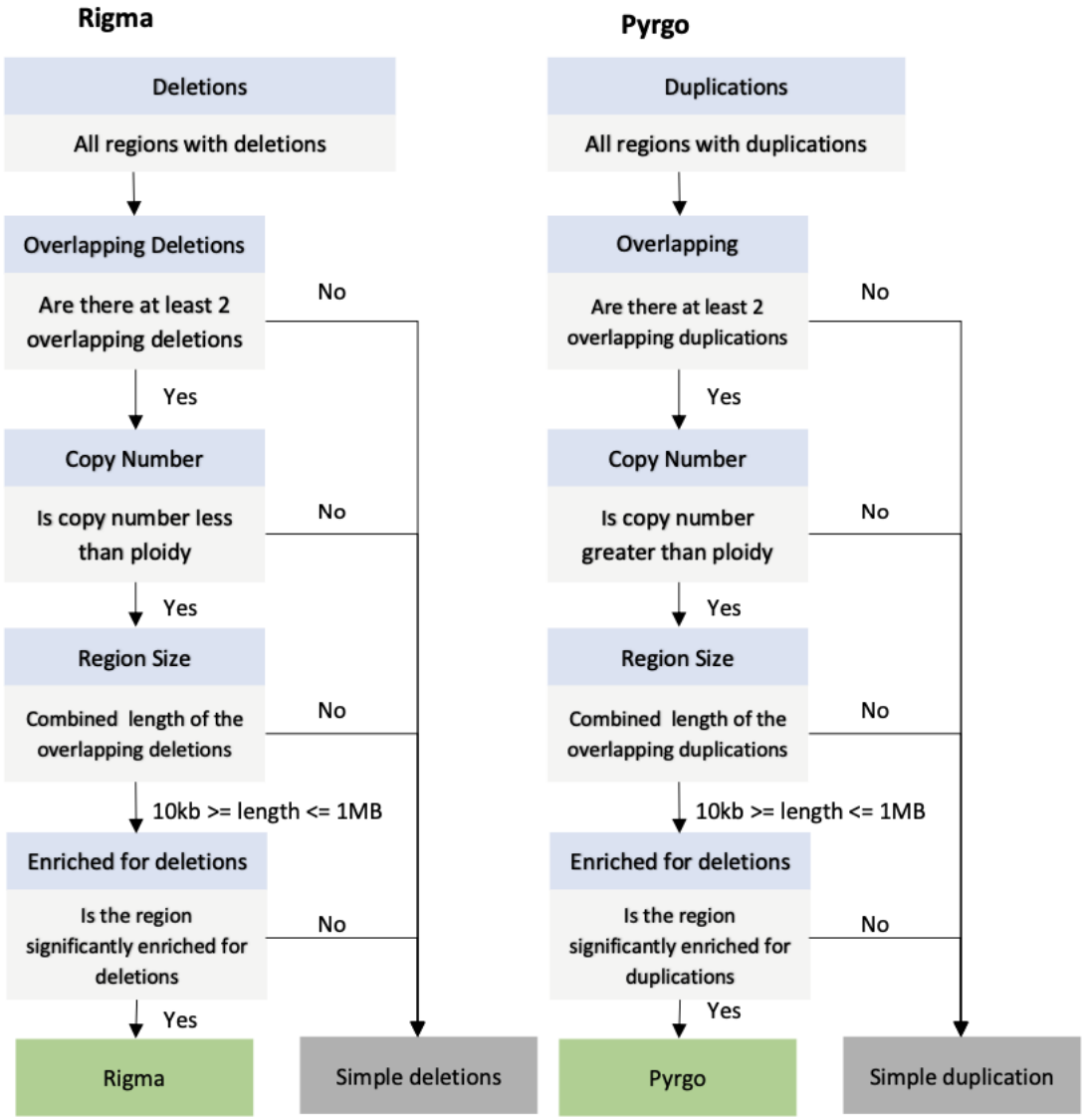


Figure 9 Flow diagram for identification of Rigma and Pyrgo by JaBbA gGenomes

This flow diagram demonstrated how gGenomes identifies regions of rigma and pyrgo based on copy number and size, location and enrichment of deletions or duplications respectively (Hadi et al. 2020).

Tyfonas, Breakage Fusion Bridges and Double Minutes

The JaBbA gGenomes pathway separates amplified regions into four classifications tyfonas, breakage fusion bridges and double minutes and other this is shown in Figure 10 (Hadi et al. 2020). Amplified regions are defined as regions with copy number of at least 7 and greater than the ploidy of the sample. Amplified candidate regions are then split based on the proportion of SVs in the candidate region that are fold back inversions

also called inverted duplications. Candidate regions with atleast 50% fold-back inversion SVs are then subdivided into tyfonas or breakage fusion bridges based on the number of junctions that have a copy number greater than 7. If atleast 26 junctions in the candidate region have a copy number >7 then the candidate region is classified as tyfonas. If the candidate region has less than 26 junction with copy number > 7 then the candidate region is classified as breakage fusion bridges.

Alternatively candidate regions with less than 50% fold back inversions are subdivided based on the number of junction with copy number > 7. To be classified as a double minute, a candidate region must have at least 31 junctions with copy number greater than 7. If the candidate regions do not meet this threshold then it is classified as other amplification.

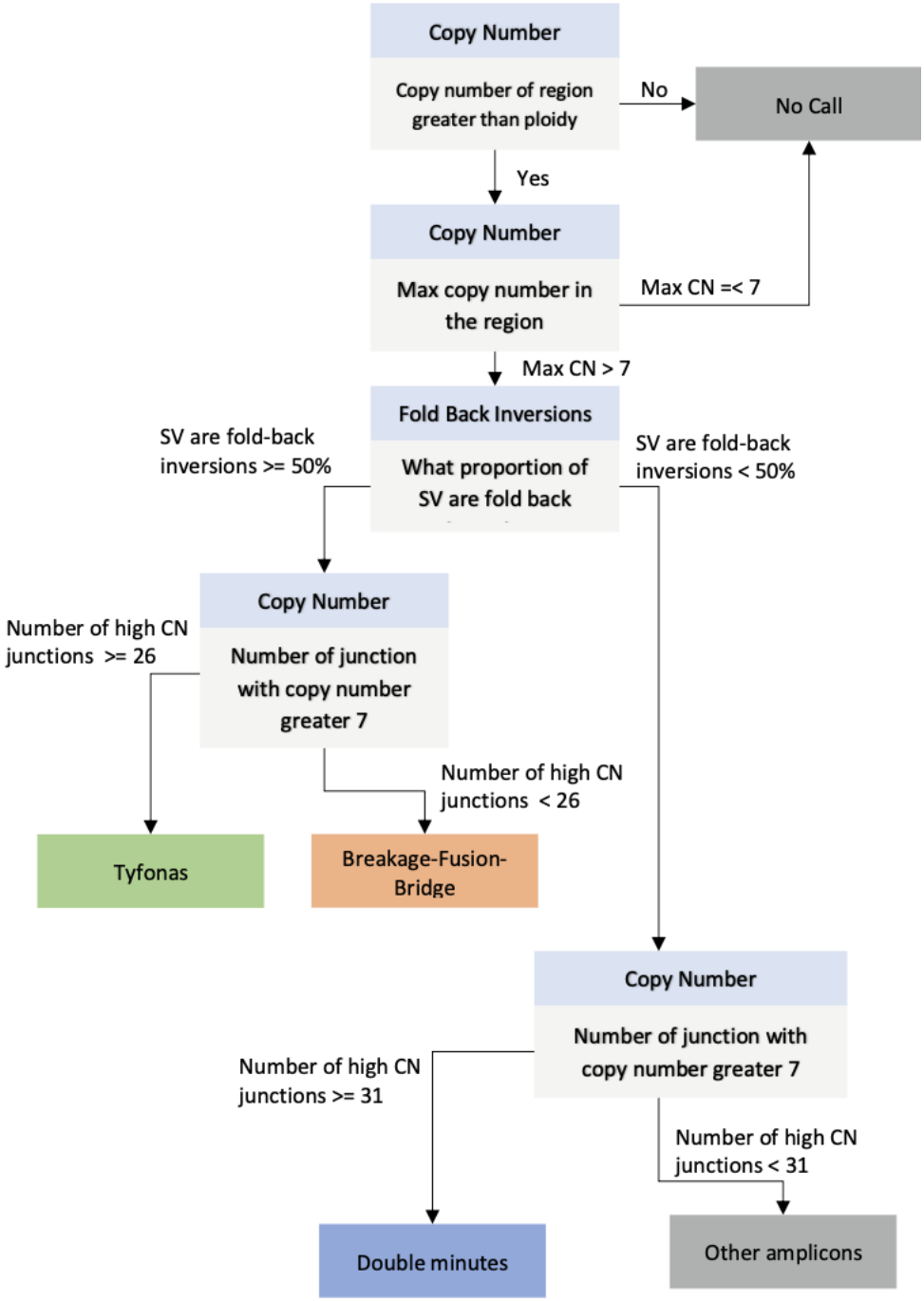


Figure 10 Flow diagram for classification of amplified region by gGenomes

The flow diagram demonstrates how gGenomes classifies amplified regions identified by JaBbA into five categories bases on copy number and proportion of fold-back inversion. The five possible classifications are Breakage Fusion Bridge, Tyfonas, double minutes, other amplicons and no call (Hadi et al. 2020).

Chromoplexy

The JaBbA gGenome classifies chromoplexy from junctions with a copy number < 3. Breakpoints within 10kb are grouped into a cluster and at least three clusters must be linked by intrachromosomal SV or SV linking cluster at least 10Mb apart to be classified as chromoplexy (Figure 11) (Hadi et al. 2020).

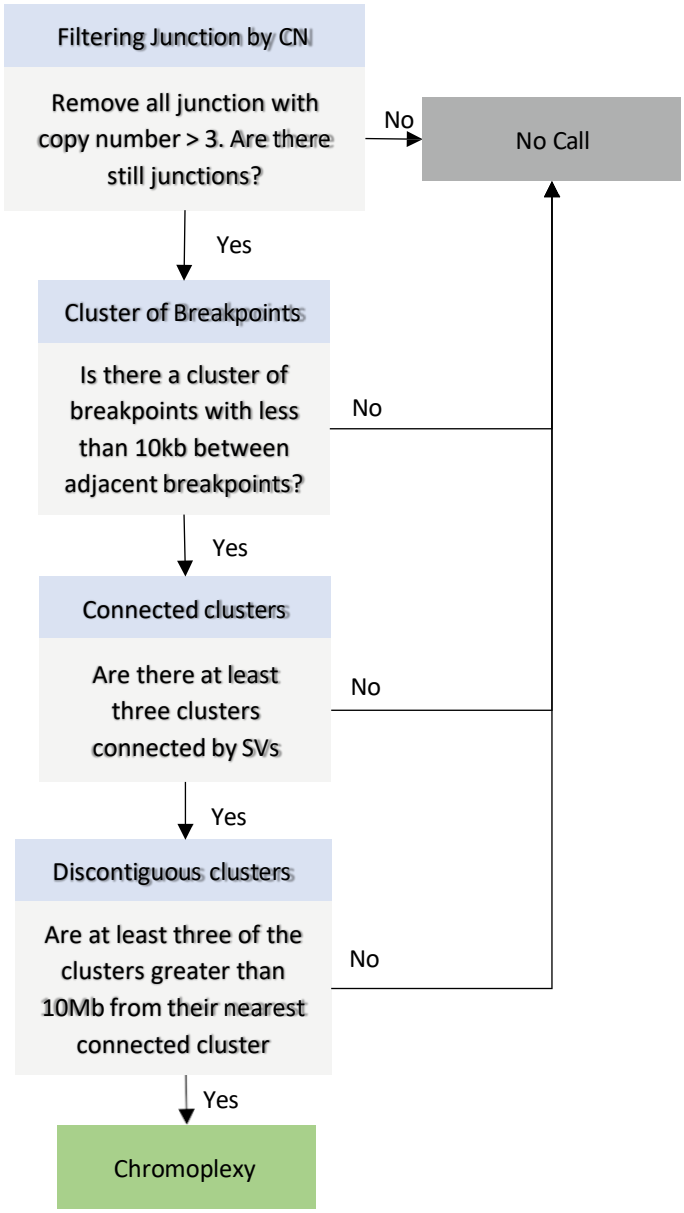


Figure 11 Flow diagram for identification of Chromoplexy by gGenomes

The flow diagram demonstrates how gGenomes identifies chromoplexy from number of SV and their clustering (Hadi et al. 2020).

The main difference between gGenome calls of chromoplexy and the criteria proposed by Baca *et al* when initially defining chromoplexy is the restriction of copy number to < 3 in the gGenome criteria. Both use discontinuous regions linked by SVs to identify chains of translocations (Baca et al. 2013; Hadi et al. 2020)

Seismic amplifications

Seismic amplifications is defined from copy number and SV thresholds (Rosswog et al. 2021). From the consensus SV and CNV calls, seismic amplifications was called using the criteria proposed by Rosswog. The criteria and thresholds used are shown in Figure 12 (Rosswog et al. 2021). From a amplified CN segments defined as copy number at least 5 or 9 if ploidy is greater than 3. The SV within the CN segment +- 5kb are then identified and amplified copy number segments linked by structural SVs are combined and if there is at least 14 SVs in the combined segments then the region is identified at seismic amplifications. If the segment has < 14 SV then it is classified as other amplicon (Rosswog et al. 2021).

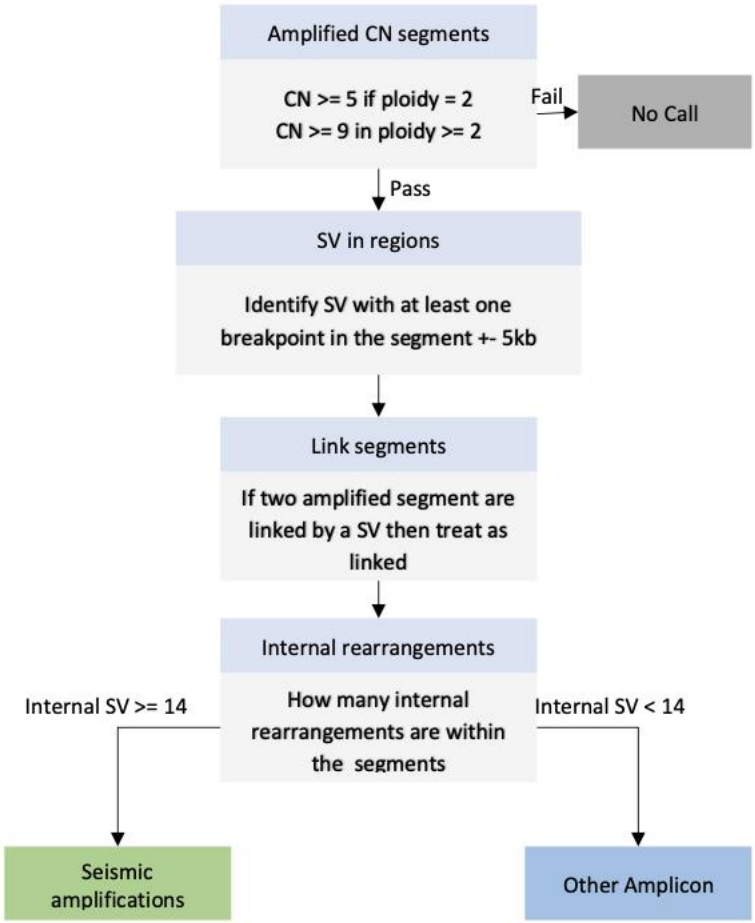


Figure 12 Flow diagram for classification of seismic amplifications by gGenomes

The flow diagram shows how the thresholds described by Rosswog et al can be used to identify seismic amplification from CNV and SV calls (Rosswog et al. 2021).

Identifying Other genomic features

Whole genome duplication

Whole genome duplication (WGD) had previously been called in the cohort by other members of the Semple Lab using the Facets package to estimate the proportion of the genome which is duplicated (R. Shen and Seshan 2016). A Facets score of at least 0.75 was called as WGD (R. Shen and Seshan 2016).

Homologous recombination deficiency

Homologous recombination deficiency (HRD) was identified by other members of the Ewing lab by implementing HRDetect method outlined in Davies et al. 2017.

Genomic instability and cSV

To investigate the relationship between genomic instability and cSV, the sample wide number and size of SV and CNV were used as a measures of genomic instability. Using a logistic regression model to investigate the impact of increased genomic instability on the risk of cSV being called. The SV or CNV measurement was centred and scaled so that the odds ratio of the logistic regression relates to an increase of one standard deviation of the measurement.

Uneven distribution of cSV across the subcohorts

The combined cohort is made up of sub cohorts of varying sizes to see if any cSV were more common in a sub cohort relative to its size than would be expected by chance and chi sq test was used.

Mutational signatures and cSV

To investigate what mutational processes may be driving the generation of cSV, the Starfish algorithm was utilised to classify samples based on Starfish mutational signatures (Bao et al. 2022). Starfish classifies samples based on copy number and breakpoint information into one of six classifications: micronuclei, double minute and ecDNA, chromatin bridges, large loss, large gain and hourglass. The first three

classifications are reported to be biologically meaningful mechanism of genomic instability (Bao et al. 2022).

Co-occurrence of cSV across the combined cohort

To investigate co-occurrence and mutual exclusivity of cSV the pairwise enrichment between cSV was assessed using the Select algorithm (Mina et al. 2017, 2020). Select used predicated null distributions bases on the prevalence of the event in this case cSV in the population being tested. To assess the magnitude and significance of the enrichment or depletion of pairwise cSV interaction a chi squared test was used.

To investigate if the different cSV types were co-occuring more than expected by chance, pairwise interaction and hierarchical clustering of the cSV was performed. The significance of these clusters was tested by bootstrapping 10,000 times to test the robustness of the clustering.

Explaining SV occurrence by cSV

To investigate what proportion of the genomic instability could be explained by the investigated cSV, the overlap of each cSV type and SV was found and the number of SV identified, which could then be compared to the total amount of SV in the samples.

In the PCAWG analysis of over 2500 tumours, clusters of SV where identified based on breakpoints being closer than expected by chance (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisation/Ins et al. 2020). Clusters of SV were identified in the combined cohort by other members of the Ewing Lab. For a cluster of SVs to be explained by a cSV, it is necessary that the cSV overlaps with at least 50% of the SV breakpoints within the cluster.

Chromosome enrichment for cSV

A chi-square test was used to examine the enrichment of each chromosome for each cSV. This involved comparing the number of observed occurrences of cSV on each chromosome to the number of expected occurrences, which was determined based on the fraction of the genome encompassed by that chromosome. Furthermore, the

number of trials in this comparison was the total number of occurrences of cSV across all chromosomes.

Fragile sites

The location of fragile sites and presence of genes in fragile sites was based on the curated information in the HumCFS database (Kumar et al. 2019). The HumCFS database combines literatures searches and manual curation to identify fragile sites in the human genome (Kumar et al. 2019).

Known oncogenes, cancer gene fusions and tumour suppressor genes

The genes listed in the Cancer Gene Census as tier 1 or 2 were used to identify oncogenes in the gene list created in this work (Sondka et al. 2018). Tier 1 genes have documented activity and evidence of mutations relevant to cancer and have the strongest evidence as oncogenes (Sondka et al. 2018). Tier 2 gene have strong evidence for being oncogenes but not as strong as tier 1 genes (Sondka et al. 2018). This list was then subdivided based on the genes annotated function as oncogenes, fusion genes and tumour suppressor genes.

Clinically Relevant Gene List

A list of genes thought to be highly clinically relevant was manually curated by other members of the Scottish HGSOC collaboration (Hollis et al. 2022). This also included genes predicted to be important from the ARIEL studies on non-BRCA HRD (da Cunha Colombo Bonadio et al. 2018).

Essential Genes

The CoRe package provides a curated list of essential genes identified from multiple CRISPR-Cas9 screens (Vinceti et al. 2021).

cSV Gene List and disrupted Gene List

To identify the genes impacted by cSV, the overlap between the genomic position of cSVs and the genomic position of all protein coding genes was found using the GenomicRanges package for R (Lawrence et al. 2013).

To identify the genes that are most likely to be affected by the presence of cSV, the gene list was first narrowed down to genes located within a cSV region that contained at least one consensus SV breakpoint within the gene. However, as genes that are not disrupted by SVs are also of interest, a second gene list was created, the enveloped genes. This new list includes genes that are completely covered by a cSV and do not have any SV starting or ending within the gene. Some genes are never covered by any cSV in the combined cohort and these genes form the elusive gene list.

Chapter 3: Prevalence and Co-occurrence of Complex Structural Variants

Introduction

This chapter will investigate the patterns of complex structural variants (cSVs) in HGSOC by considering their prevalence and co-occurrence within and across cohorts. Additionally, the interdependencies between cSVs and simple structural variants along with other relevant genomic features such as homologous recombination deficiency (HRD) and whole genome duplication (WGD) will be explored.

The prevalence of simple structural variants (SVs) varies widely between cancer types, with fewer than ten structural variants identified in myeloproliferative neoplasms samples and more than a thousand in soft tissue liposarcoma samples (Y. Li et al. 2020). In the PCAWG study of 2559 cancer whole genomes from 38 tumour types, one or more structural variants were found in 2429 (95%) samples with deletions being the most common type of SV identified (Y. Li et al. 2020). HGSOC samples in PCAWG were highlighted as particularly structurally complex with high numbers of tandem duplications, deletions and unbalanced translocations, as well as abundant but poorly described 'complex unclassified' variants (Y. Li et al. 2020).

It is important to note that the accurate identification of structural variants remains an ongoing challenge in computational biology and can result in different structural variants being identified from the same whole genome sequencing (WGS) data. To create the highest confidence set of SVs, this work will utilise consensus SV calls. The consensus calls will allow for comparisons to be made between cSV types that require SVs called by different SV callers.

Identifying cSVs based on patterns of SVs is unavoidably challenging since identifying SVs is an ongoing challenge in itself. However, this is compounded by the caveat that the identification of the patterns associated with each cSV type is an additional ongoing challenge. For example, ecDNA have estimates of prevalence ranging from 85% to 14%

of cancer samples (Cen et al. 2022; Zeng, Wan, and Wu 2020). The range of frequencies observed in the literature highlights perhaps the most important caveat in the field of cSV research. The criteria and method used to identify the cSV has a massive impact on the cSVs identified, as any change in criteria and method will result in sampling from a different point on a spectrum of complexity.

The combined cohort is the largest uniformly processed cohort of HGSOC whole genome sequenced samples and uses several complex structural variant callers to identify cSVs, which is nearly three times the number of HGSOC samples used in the PCAWG study (Y. Li et al. 2020). The increased size of the cohort should allow for more comprehensive detection of cSVs, including rarer variants.

In the literature, different criteria for identifying cSV have frequently been used. For example, chromothripsis has been predicted simply from the presence of relatively few SVs identified on a chromosome (Molenaar et al. 2012; Northcott et al. 2012; Rausch et al. 2012; Chiang et al. 2012). Alternatively, more sophisticated patterns based on combinations of SVs and CNVs to determine chromothripsis have also been used (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). In order to have consistent high quality calls of cSV, this work is based upon a variety of recently published algorithms (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020).

In the literature the analysis of cSV often only focuses on one or a few types of cSVs. In this work, 8 cSV types have been identified within the same HGSOC samples (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020; Cen et al. 2022). The breadth of cSV types identified allows for the investigation of co-occurrence and mutual exclusivity between cSV types. The combined cohort has also been profiled using RNAseq and annotated for the presence of genomic features such as HRD and WGD by other members of the Semple and Ewing labs, allowing the relationship between cSV and other features to be investigated.

Recent advances in the treatment of HGSOC focus on samples that are HRD, these

treatment options based on PARP inhibitors have shown promising results in improving patient survival (da Cunha Colombo Bonadio et al. 2018). However, 44% of HGSOC samples in the combined cohort are not HRD and therefore cannot benefit from these improved treatments. The inability of tumours with HRD to repair double strand breaks (DSBs) using the homologous recombination repair pathway leads to structural variation within the tumour cell population, which may be subject to selection during tumorigenesis (Talseth-Palmer and Scott 2011; Pennington et al. 2014). However, the HGSOC samples that do not have HRD are also structurally diverse and may exploit other sources of genomic variation, such as the cSVs explored in this chapter.

The key questions addressed in this chapter are:

i. Prevalence: What is the prevalence of different complex structural variant types in the combined cohort, and does this vary across sub-cohorts?

ii. Co-occurrence: Are complex structural variant types independent of each other, or do they show co-occurrence and mutual exclusivity?

iii. Genomic features: What is the relationship between complex structural variants and other genomic features, such as whole genome duplication and homologous recombination repair deficiency?

iv. Proportion: What proportion of genomic complexity can be explained by known complex structural variants?

v. Novel complex structural variants: Is there evidence for the presence of novel complex structural variants based on structural variant / copy number variant patterns across samples?

High Grade Serous Ovarian Cancer is Highly Genomically Unstable and this is Consistent Across the Combined Cohort

HGSOC is a highly genomically unstable type of cancer where there are hundreds to thousands of SVs and CNVs spread across a single tumour genome (Morden et al. 2021). The combined cohort of HGSOC examined here has been uniformly analysed but consists of five sub-cohorts which initially were sampled and sequenced at different times and from different groups of patients, potentially resulting in technical and biological variability. Any large differences in the landscapes of structural variants and copy number variants across the genome between sub-cohorts could then be propagated into differences in the prevalence and co-occurrence of cSVs. On the other hand, consistent patterns seen across sub-cohorts are likely to represent common features of HGSOC biology.

SV and CNV were identified using multiple callers by other members of the Semple and Ewing Labs who also generated consensus calls for both SV and CNV. It is these consensus calls that are discussed below. To identify a high confidence set of structural variant calls, a consensus was taken from between GRIDSS and Manta callers by other members of the Ewing lab (Chen et al. 2016; Cameron et al. 2017; Cameron, Baber, Shale, Papenfuss, et al. 2019). An overlap of at least 100 bps between both SV calls was required for it to be called a consensus SV. A consensus was taken among copy number calls identified by CNVkit, CLImAT and PURPLE (Talevich et al. 2016; Yu et al. 2014; Cameron, Baber, Shale, Papenfuss, et al. 2019). Tacking a consensus from multiple callers is recommended to improve accuracy based on bench marking test (Coutelier et al., 2022)

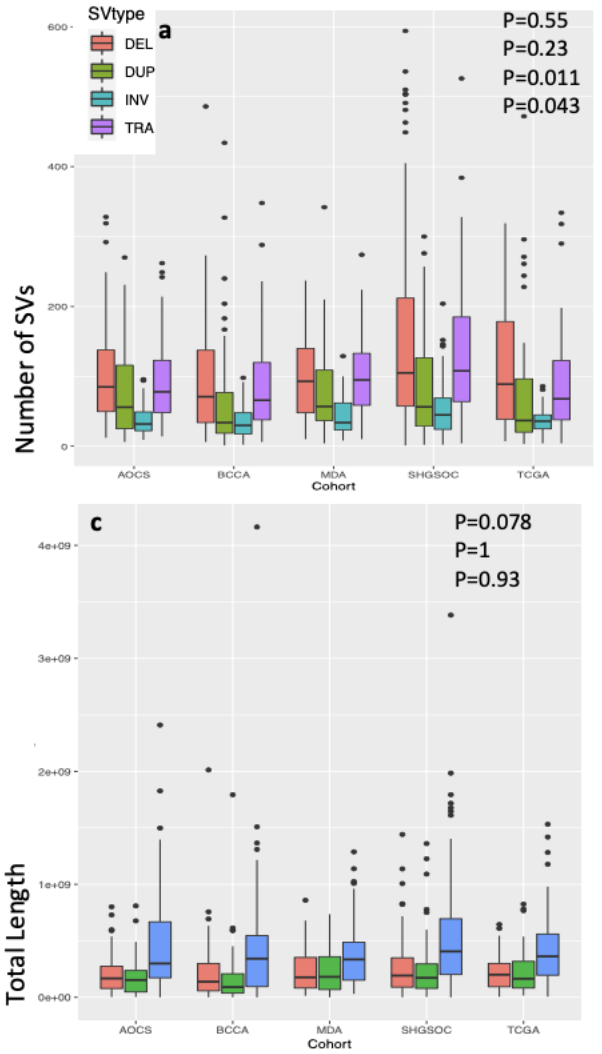
The burden of structural variant types is broadly consistent across sub-cohorts

To investigate the consistency of the burden of structural variants across sub-cohorts, we assessed the differences in the abundance of each type and the lengths of genome impacted. The total number of each SV type as well as the total length of the sample covered by each SV type is shown in Figure 13. Overall, there was consistent

correspondence between sub-cohorts, with the highest burdens observed for deletions and translocations, and notably lower frequencies of duplications and inversions in all sub-cohorts.

The SV burden by cohort was marginally different, but not significantly, for the number of translocation and inversions shown in Figure 13. However, the differences between cohorts did not reach statistical significance for the total length of sample covered by each SV type with deletions, duplications and inversions. The significant difference observed in the Kruskal-Wallis test for number of SVs in each cohort was further investigated using Dunn's post-hoc analysis to identify which sub-cohorts were contributing to the difference. This showed that the difference observed was due to the SHGSOC sub-cohort having more SV of each type than the BCCA sub-cohorts, this is shown in Figure 13 b**.**

Figure 13 also shows that across all sub-cohorts, the mean total length of inversions was greater than both deletion and duplications. Additionally, it indicates that the number of inversions identified in each cohort were the lowest of any SV meaning that that inversions are the least frequent but largest SV type. This is consistent with the functional impact of inversions being generally less severe than the impact of deletions, as expected.



b

Dunn's Post-Hoc Number of Inversions	
Cohorts	P
BCCA-SHGSOC	7.84×10^{-4}

Dunn's Post-Hoc Number of Translocation	
Cohorts	P
BCCA-SHGSOC	0.049

Figure 13 The Burden of Structural Variants is Consistent Across the Combined Cohort

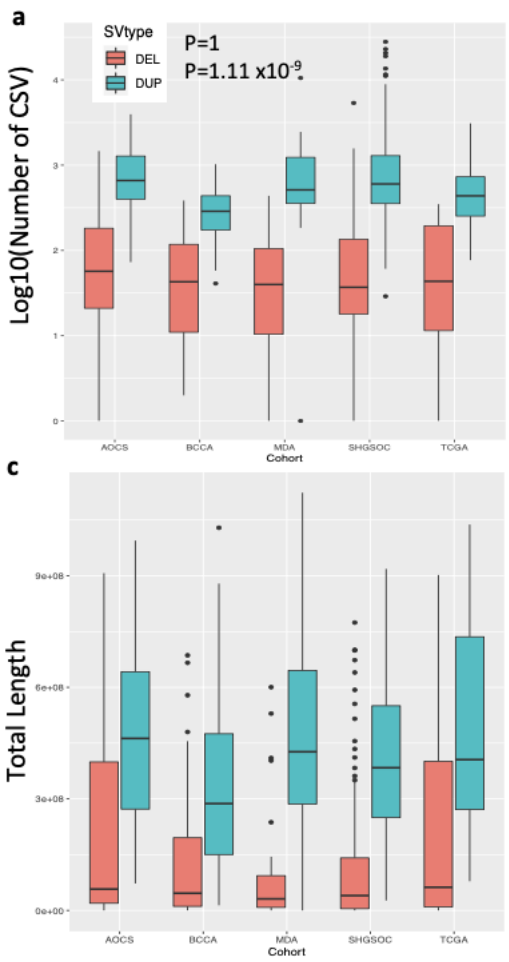
a. Shows the number of consensus deletion, duplication, inversions and translocations within a sample grouped by sub-cohort. The P values for Kruskal-Wallis test comparing the number of deletions, duplications, inversions, and translocations per cohort is shown in red, green, blue and purple respectively. For the structural variant types that showed a significant difference the Dunn's Post-Hoc analysis was performed and significantly different cohort combinations are shown in **b**. The total length of consensus deletion, duplication, and inversions within a sample grouped by sub cohort is shown in **c**. Translocations were excluded from this as they do not have a biologically meaningful length. The P values for Kruskal-Wallis test comparing the amount of the genome

impacted by deletions, duplications and, inversions, per cohort is shown in red, green and blue respectively. P values adjusted by Bonferroni correction.

Burden of Copy Number Variants is Broadly Consistent Across Sub-Cohorts

Similar to the investigation number of SVs and size of SVs between sub-cohorts, the numbers and total lengths occupied by consensus CNV types were examined using a Kruskal-Wallis test (Figure 14**)**. The number of duplications was significantly different between the sub-cohorts. However, the number of deletions was not significantly different between sub-cohorts. As can be seen from Figure 14 b**,** the difference in the numbers of duplications between sub-cohorts is driven by the BCCA cohort. The total length of consensus CNV similarly showed a statistical difference between sub-cohorts for duplication. However, the difference for deletions did not reach statistical significance. Further investigation of the statistical difference between sub-cohorts using Dunn's post-hoc analysis suggests that the BCCA cohort has fewer and smaller duplication than the other sub-cohorts. There was a consistent trend across the sub-cohorts, showing duplations to be more numerus and cover more of the genome.

Figure 13 and Figure 14 both show consistent trends across all sub-cohort, an important observation from Figure 13 and Figure 14 is that the AOCS cohort did not appear to be significantly different from the other cohort despite it being the only cohort where resistant or relapsed cases were selected for WGS. Additionally Figure 13 and Figure 14 also highlight the highly rearranged nature of HGSOC with samples containing hundreds of SV and thousands of CNVs. Figure 15 shows that CNV events are larger than SV events, and that the copy number duplications are the most common events across the combined cohort.



b

Dunn's Post-Hoc Number of Duplications	
Cohorts	P Adj
AOCS-BCCA	6.11x10-09
BCCA-MDA	7.06x10-03
BCCA-SHGSOC	1.11x10-08
AOCS-TCGA	0.041

d

Dunn's Post-Hoc Size of Genome Impacted by Duplications	
Cohorts	P Adj
AOCS-BCCA	0.0012
BCCA-TCGA	0.018

Figure 14 The Burden of Copy Number Variation is consistent across the Combined Cohort

a. Shows the number of consensus deletion and duplication, within each sub cohort. The P values for Kruskal-Wallis test comparing the number of deletions or duplications per cohort is shown in red and blue respectively. For the structural variant types that showed a significant difference, the Dunn's Post-Hoc analysis was performed, and significantly different cohort combination are shown in **b**. The total length of consensus deletion and duplication each sub cohort is shown in **c**. The P values for Kruskal-Wallis test comparing the number of deletions or duplication per cohort is shown in red and blue respectively. For the structural variant types that showed a significant difference,

the Dunn's Post-Hoc analysis was performed. The significantly different cohort combination are shown in **d**.

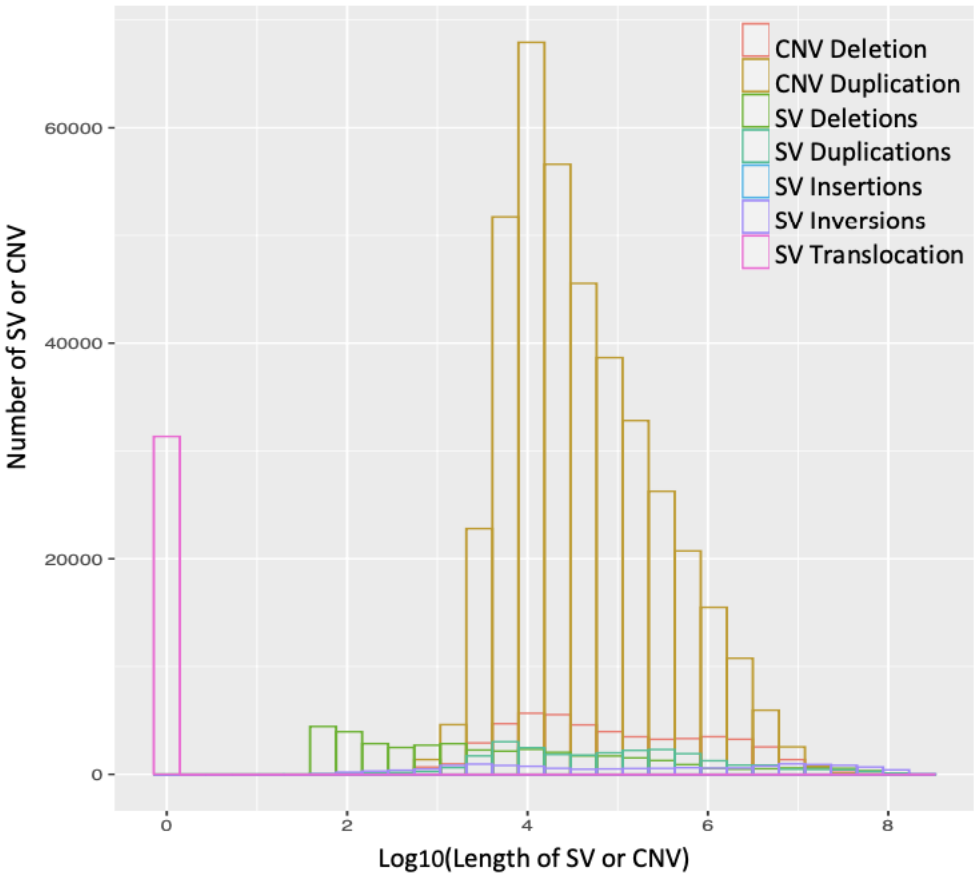


Figure 15 CNVs and SVs occur on different scales

The histogram shows the size of consensus SV and consensus CNV calls within the combined cohort.

Complex Structural Variants are abundant in High Grade Serous Ovarian Cancer

As stated previously, patterns of simple structural variants and copy number changes can be classified into different cSVs. In this section, the prevalence of eight types of complex structural variants will be assessed: chromothripsis, chromoplexy, ecDNA, breakage fusion bridge (BFB), pyrgo, rigma, tyfonas, and, seismic amplification (SA). The number of samples in which each cSV was identified in the combined cohort are presented in Table 2, together with genomic features such as WGD and HRD.

Table 2 also shows the number of papers that mention each cSV type in the title or text and shows the disparity in how much different cSV types have been studied. Some cSV types, such as pyrgo, rigma and seismic amplification have only been reported in the publication in which they were originally described whereas, chromothripsis and BFB have been widely reported in hundreds of publications. It is also likely that ecDNA are underrepresented by this measure as they have previously been referred to by alternative names such as double minutes (Ilić et al. 2022).

	Number of Samples	Percentage of the Combined Cohort	Detection Algorithm	Published Papers
Homologous recombination repair deficient	182	56%	HRDetect	2550
Chromoplexy	181	55%	JaBba-gGenome	58
Whole genome duplication	159	49%	Facets	492
Chromothripsis	107	33%	ShatterSeek	445
Pyrgo	92	28%	JaBba-gGenome	1
Breakage Fusion Bridge	86	27%	JaBba-gGenome	211
Rigma	62	19%	JaBba-gGenome	1
Extrachromosomal circular DNA	53	16%	AmpliconArchitect	191
Seismic amplifications	19	6%	Simple thresholds	1
Tyfonas	7	2%	JaBba-gGenome	2

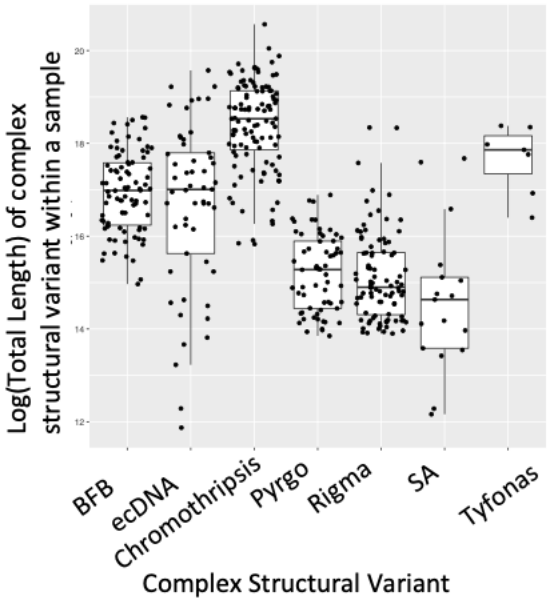
Table 2 Prevalence of complex structural variants and other genomic features

The table shows the number of samples with each complex structural variant and other genomic features as well as the algorithm used to detect it. Publications were identified from PubMed by searching the cSV type and restricted to papers that also mention cancer. The number of publications was correct as of September 2022. The cSV types were identified using the following algorithms: the methodology of HRDetect to identify homologous recombination repair deficient; ShatterSeek to identify chromothripsis (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020); JaBba-gGenome to identify chromoplexy, pyrgo, breakage fusion bridge, rigma and tyfonas (Hadi et al. 2020); Facets to identify whole genome duplication; AmpliconArchitect to identifiy extrachromosomal circular DNA (Deshpande et al. 2019); seismic amplifications was identified using the thresholds set out in (Rosswog et al. 2021).

Table 2 shows that the prevalence of cSV within the combined cohort can vary from more than half the cohort in the case of chromoplexy to two percent in the case of tyfonas. Consistent with these data (Table 2), PCAWG reported chromothripsis in roughly a third of the HGSOC samples they studied (N=110) (Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020) using the same algorithm used here (ShatterSeek).

The prevalence of 16% of samples with ecDNA in the combined cohort fits well with previously reported ranges of prevalence across cancer types. The presence of ecDNA has been reported in half of all cancer types with prevalence ranging from 1.4% to 31% (Turner et al. 2017). Glioblastoma has been reported to have ecDNA in 60% of samples (Ilić et al. 2022; X.-K. Zhao et al. 2021a). Work 56ormalize the PCAWG data set reported that pan cancer, 14.3% of samples have ecDNA (Zeng, Wan, and Wu 2020).

The genomic span of each cSV type and their overlap with simple SV and CNV alterations is shown in Figure 16 – Figure 18. These data examined cSVs at the sample level, since a sample may have multiple calls of a cSV type. ShatterSeek the algorithm used for detecting chromothripsis can only define a region within a chromosome meaning that two regions of chromothripsis on separate chromosomes even if linked by multiple translocations will be called as separate chromothripsis regions. However, the callers based upon graph-based methods (JaBba-gGenome and AmpliconArchitect) can link regions from multiple chromosomes via SVs and therefore, a sample level description of the cSV is the most comparable between cSV types.



cSV	Median	IQR
ecDNA	24322947	47415501
BFB	23797534	31903072
Chromothripsis	111630558	146897188
Prygo	2955229	4607783
Rigma	4324288	6112962
Seismic Amplification	2268044	2868799
Tyfonas	57110564	41593610

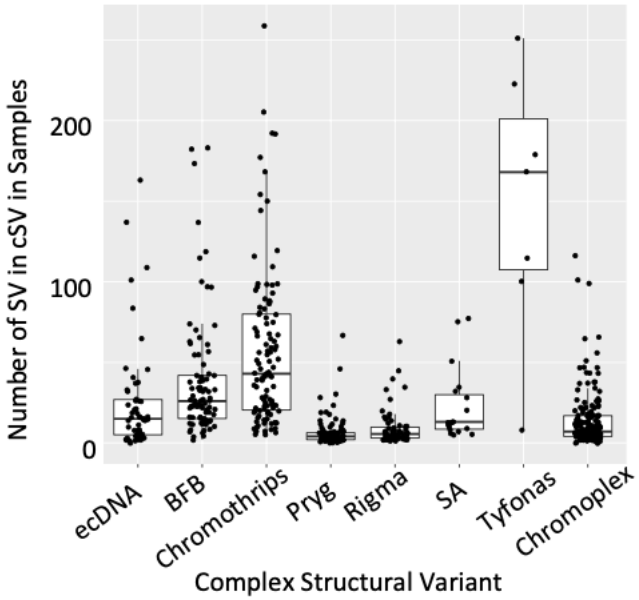
Figure 16 Complex structural variants impact large genomic regions

The total length of the genome impacted by each complex structural variants (cSV) type across samples impacted by that type. The median and interquartile range for each cSV type is shown in the table.

Figure 16 shows the size of the region impacted by cSV within a sample where the cSV is present. The largest event shown was chromothripsis, with a median size of over a 100 MB across the genome, is consistent with other methods of identifying chromothripsis in the literature where a single region of chromothripsis could be identified using 50Mb windows (Voronina et al. 2020). The large multi MB size of the interquartile ranges of all of the cSV types indicate substantial variation in the amount of the genome impacted by cSV when present, this underscores the heterogeneity of cSV of the same types in different samples. The general trend is that chromothripsis, BFB, ecDNA and tyfonas are

larger events with ecDNA having the large range of sizes while pyrgo, rigma and seismic amplification are smaller more focal events. However, it should be noted that the median size of the genome impacted by all the complex structural variant types was greater than 2Mbs. This highlights the large-scale rearrangements described by these events. Chromoplexy was excluded from this analysis as it is a chain of reciprocal translocation breakpoints assigned a short length by the caller, and therefore length is not a biologically meaningful measure.

Another measure of the scale of cSVs is the number of simple SVs within the cSV, based upon the consensus SV calls for all SV types (duplication, deletions, inversion and translocations) within each cSV region (Figure 17).



cSV	Median	IQR
ecDNA	9.5	14.5
BFB	17	29.25
Chromothripsis	19	16
Pyrgo	3	4
Rigma	5.5	6.75
Seismic Amplification	13	21.5
Tyfonas	168	93.5
Chromoplexy	7	13

Figure 17 Number of SV in cSV across the combined cohort

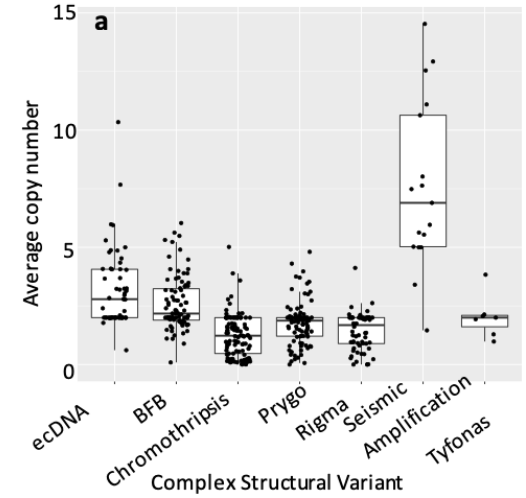
The total number of consensus SV within a cSV region for each sample impacted by that cSV. A sample can have multiple calls of the same cSV across the genome. Additionally, the median and interquartile range of the number of SV within a cSV is also shown.

Tyfonas has the highest median number of SV of any cSV but, as it is only present in 7 samples and it has a large interquartile range, this could be due to small sample size. The number of SV and CNV in samples is often reported vaguely in the literature by describing events such as chromothripsis as containing tens to hundreds of SVs and many changes in copy number(Maher and Wilson 2013; Shorokhova, Nikolsky, and Grinchuk 2021). Within the combined cohort, the median number of SV in

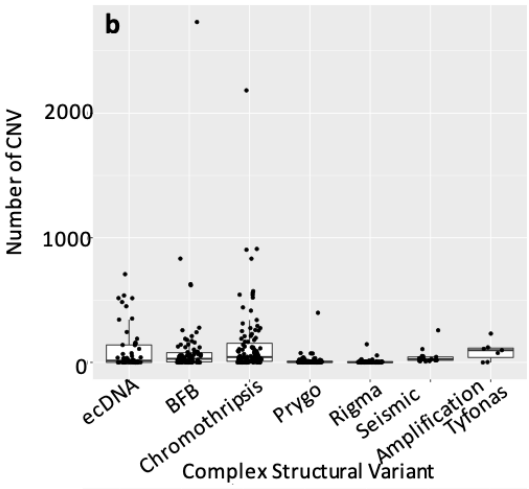
chromothripsis was 19 with 14 samples containing more than 100 consensus SVs meaning the chromothripsis regions identified are consistent with the range suggested by the literature (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020; Korbel and Campbell 2013). Overall, chromothripsis and BFB types account for the highest numbers of simple SV calls.

Complex structural variants are not just combinations of SVs but frequently involve alterations to copy number. This means an alternative view of cSV can be defined by three measures: 1) the number of CNV calls within a cSV region, 2) the average copy number across a cSV region and 3) the maximum copy number within the cSV region.

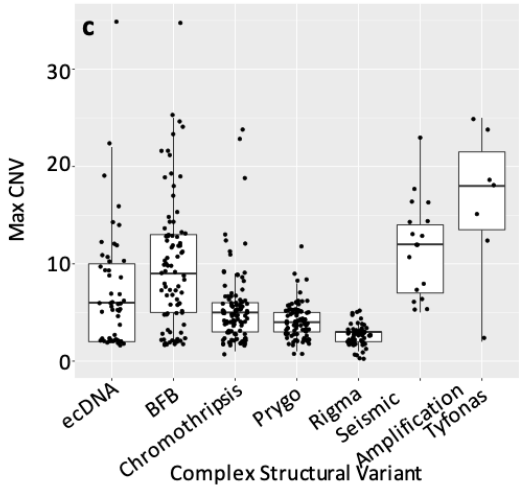
In Figure 18 a, the average copy number of a cSV per sample in which the cSV is present is displayed for all cSV types, including those known to involve amplifications, such as ecDNA, BFB, seismic amplification and tyfonas. It is clear that high copy number amplifications are generated disproportionately by the seismic amplification, ecDNA and BFB types (Figure 18 a). This tendency is also reflected in the maximum copy number estimates for the same cSV types (Figure 18 c). This suggests that tumours acquiring amplifications of particular genes or regions will be provided with more DNA by those cSV types.



cSV	Median	IQR
ecDNA	2.60	2.03
BFB	2.45	0.99
Chromothripsis	1.23	1.53
Pyrgo	2.00	1.36
Rigma	1.97	1.63
Seismic Amplification	7.21	5.11
Tyfonas	2.09	0.62



cSV	Median	IQR
ecDNA	1	3
BFB	2	10
Chromothripsis	43	131
Pyrgo	2	4
Rigma	1	2
Seismic Amplification	3	1
Tyfonas	2	4



cSV	Median	IQR
ecDNA	6	4
BFB	9	5
Chromothripsis	5	3
Pyrgo	3	2
Rigma	2	1
Seismic Amplification	12	7
Tyfonas	18	6.5

Figure 18 CNV in cSV events across the combined cohort

a The average copy number across the region identified as having a cSV within a sample. **B** The number of CNV calls. **C** The maximum copy number within all the regions identified as having a cSV within a sample. The median and interquartile range is shown for each copy number measure in the adjacent table.

Complex Structural Variant Types and General Measures of Genomic Instability

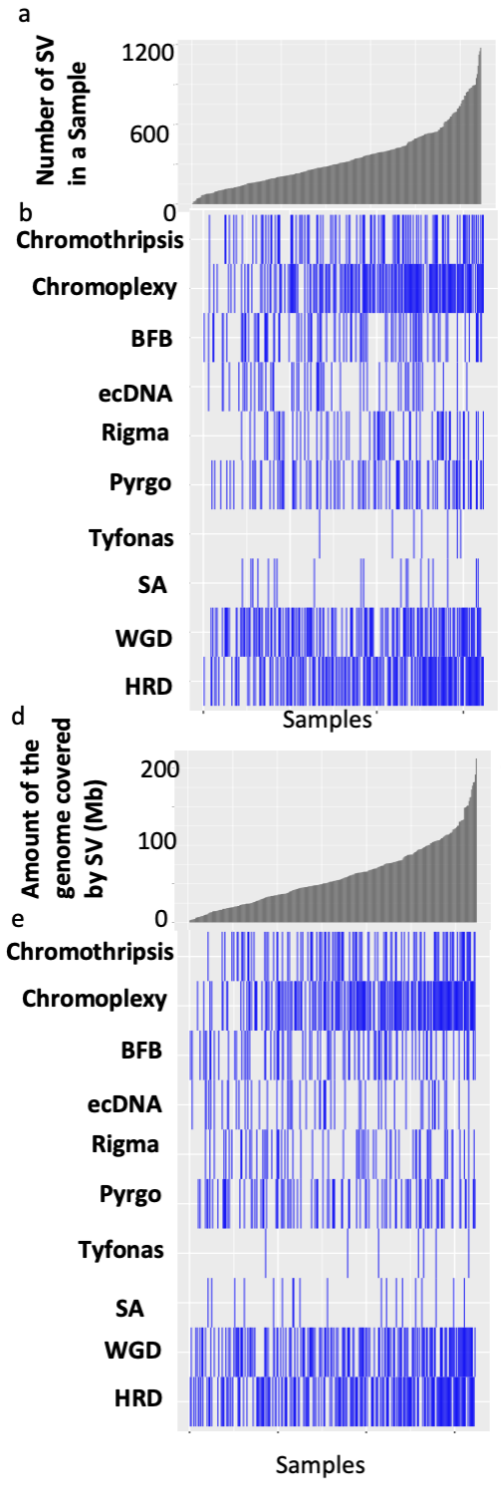
As cSVs are made up of SVs and CNVs, it may be that, cSVs are more likely to be identified within the most genomically unstable samples containing high numbers of SV and CNV calls. In Figure 19, I explore two genomic instability measures and their influence on cSV occurrence using logistic regression. Those measures are the total number of SVs genome-wide predicted per sample in the combined cohort (Figure 19 a, b, c), and the total length of the genome encompassed by those SVs per sample (Figure 19 d, e, f). The odds ratios represent the change in the probability of observing a given cSV type for a sample as genomic instability increases. For some cSVs, there is a relationship between genomic instability and the occurrence of cSV, such as chromothripsis and chromoplexy (Figure 19 a, b, c). However, for other cSV types such as BFB, genomic instability appears to be a very poor predictor of cSV, suggesting that cSV types differ in their relationship to simple measures of genomic instability.

Figure 19 shows that, like chromothripsis and chromoplexy, WGD and HRD are also identified more often than expected by chance in the most genomically unstable samples. This is seen consistently across both measures of instability (Figure 19), and is consistent with the existing literature. HRD is known to increase genomic instability in tumours (Frey and Pothuri 2017; den Brok et al. 2017; Hoppe et al. 2018) and WGD has been shown to rapidly induce genomic instability, within the first cell cycle following tetraploidisation (Gemble et al. 2022).

In Figure 19**,** genomic instability was measured in terms of genome-wide SV burden and compared with cSV occurrence. However, cSV calls are dependent on the presence of abundant SV calls, potentially creating spurious associations between cSV and SV frequencies. To account for this potential bias the analysis was repeated using SV numbers per sample, but excluding the SVs directly associated with the cSV identified in each sample.

Figure 20 presents a similar logistic regression analysis of the associations between

genomic instability and cSV occurrence, but for measures of instability based upon genome-wide CNV rather than SV burdens in the combined cohort. For genomic instability as measured by CNV burden (Figure 20) the only significant association found was between the total number of CNV and the occurrence of chromothripsis. This may be because, as previously mentioned, chromothripsis is the only cSV type to have a minimum number of copy number changes as one of its requirements to be identified.



c

P Value	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
1.18x10-3	1.44	1.15	1.84
5.26x10-10	2.94	2.17	4.14
1	1.01	0.78	1.28
1	0.72	0.50	0.99
0.27	1.34	1.03	1.74
0.11	1.36	1.07	1.72
0.40	1.87	1.00	3.40
1	1.19	0.75	1.80
2.64x10-4	1.33	1.07	1.68
3.42x10-8	2.45	1.84	3.34

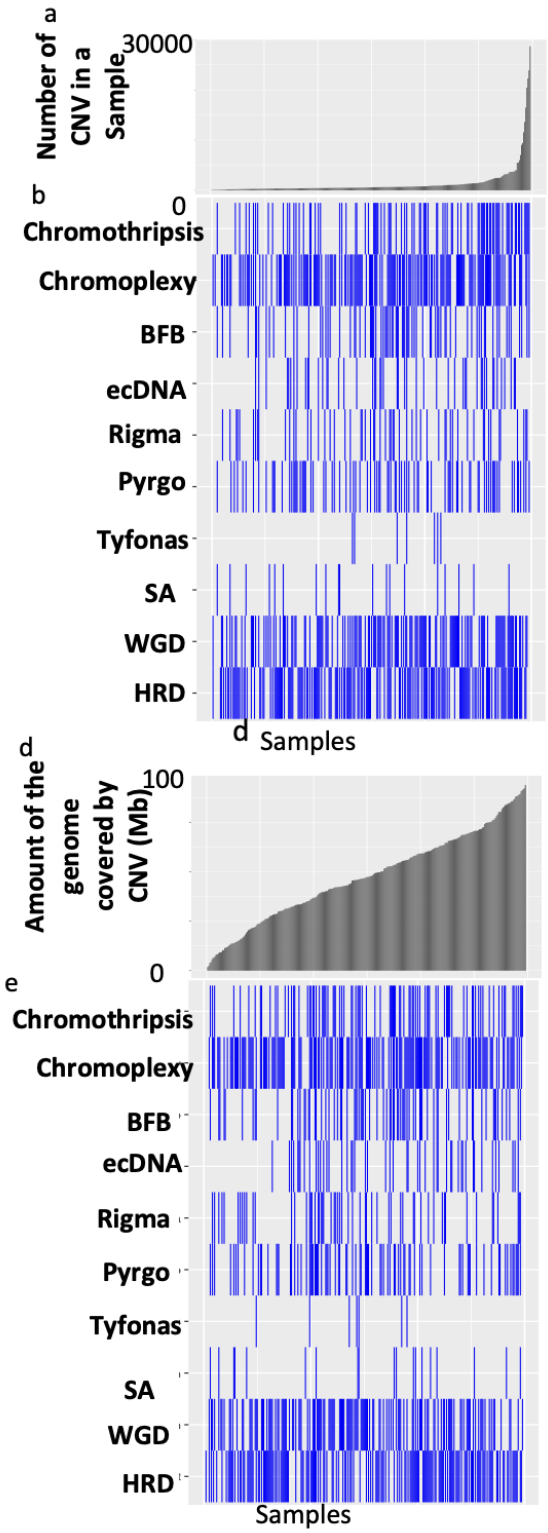
f

P Value	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
$1.18x10^{-3}$	1.60	1.26	1.84
$2.88x10^{-12}$	3.50	2.54	4.14
1	1.11	0.87	1.28
1	0.85	0.61	0.99
1	1.12	0.85	1.74
1	1.18	0.93	1.72
0.66	1.78	0.93	3.40
1	1.06	0.65	1.80
$2.64x10^{-4}$	1.68	1.33	1.68
$5.48x10^{-4}$	1.67	1.31	3.34

Figure 19 Chromothripsis and chromoplexy are associated with increased genomic instability measured by structural variant burden

a and **b** show the number and amount of the genome covered by structural variant (SV) in samples in ascending order respectively and **e** show the presence of a cSV in a sample.

The order of samples is the same as in **a** and **d**. **c** and **f** show the results of a logistic regression model for the effect of genomic instability as measured by numbers of SV and size of genome covered by SV respectively. For the logistic regression, the measurement of genomic instability was centred and scaled. P values corrected for multiple testing using the Bonferroni adjustment.



c

P Value	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
1.81×10^{-2}	1.65	1.26	2.35
1	1	0.85	1.20
1	1.09	0.91	1.29
1	0.90	0.60	1.13
1	1.09	0.90	1.30
1	1.11	0.94	1.32
1	0.58	0.04	1.27
1	0.64	0.16	1.15
1	1.02	0.87	1.22
1	0.93	0.79	1.11

f

P Value	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
0.62	1.18	0.99	1.40
1	1.01	0.85	1.19
1	1.09	0.91	1.31
1	1.18	0.95	1.45
1	0.92	0.74	1.13
1	1.03	0.87	1.24
1	0.91	0.49	1.53
1	1.02	0.72	1.41
1	0.95	0.81	1.12
1	1.07	0.91	1.26

Figure 20 Chromothripsis is associated with increased genomic instability measured by number of copy number variants

a and **e** show the number and amount of the genome covered by copy number variant (CNV) in samples in ascending order respectively and **e** show the presence of a cSV in a sample. The order of samples is the same as in **a** and **d**. **c** and **f** show the result of a logistic regression modelling the effect of genomic instability as measured by number of SV and size of genome covered by SV respectively. For the logistic regression, the measurement of genomic instability was centred and scaled. P values corrected for multiple testing using the Bonferroni adjustment.

The work presented in Figure 19 assumes that the number of SVs present in a sample is independent of the presence of a cSV. However, since the prediction of a cSV depends upon SVs, this assumption is not true.

Complex Structural Variant	P Value	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
Chromothripsis	0.18	1.32	1.05	1.67
Chromoplexy	$1.26x10^{-08}$	2.55	1.91	3.49
BFB	1	0.85	0.65	1.10
ecDNA	0.10	0.62	0.42	0.87
Rigma	0.53	1.29	0.99	1.68
Pyrgo	0.21	1.32	1.04	1.67
Tyfonas	1	1.28	0.61	2.38
SA	1	1.08	0.67	1.66
WGD	$2.64x10^{-04}$	1.33	1.07	1.68
HRD	$3.42x10^{-08}$	2.45	1.84	3.34

Table 3 Chromoplexy is significantly associated with increased genomic instability

The results of a logistic regression relating the presence of each cSV type to the number of SVs in the sample excluding those explained by the cSV in the sample. The number of SV present in the samples was centred and scaled so that the odds ratio represents an increase in one standard deviation of the amount of SV in a sample. P values were adjusted using the Bonferroni correction.

In the revised analysis (Table 3) which removes the SV that are part of the CSV from the analysis the only significant association found between genomic instability and cSV occurrence is for chromoplexy, which was more than twice as likely in samples with SV frequencies more than one standard deviation above the mean. The association of genomic instability with HRD and WGD also remains significant. In contrast, the original association with chromothripsis is weaker and no longer significant, suggesting some bias in the original association observed (Figure 19). This is consistent with the high numbers of SV per sample often found to be associated with a region of chromothripsis (Figure 17).

Sub-cohort differences in cSV Occurrence

As previously mentioned, the combined cohort is based upon sub-cohorts generated by different researchers and based upon different groups of patients. For example, the AOCS sub-cohort was selected after patients progression that did not respond to treatment. It is therefore of interest to explore the differences in cSV occurrence between sub-cohorts. Previously we observed that the distributions of SV and CSV are broadly consistent across all sub- cohorts with only the BCCA sub-cohort showing some depletion for certain SV and CNV types when compared with the other sub-cohorts (Figure 13 and Figure 14). The same sub-cohort also appears to be somewhat depleted for cSV, and particularly chromothripsis. This is consistent with the fact that cSV predictions are composed of SV and CNV (Figure 21).

However, there is also a significant enrichment of ecDNA within the AOCS sub-cohort, with ecDNA observed almost 9 times more often than would be expected by chance. This is an intriguing result given the reported association of ecDNA with the evolution of treatment resistance in multiple cancer types including ovarian cancer (Ashique et al. 2022).

The reason for the lack of chromothripsis depletion in the BCCA sub-cohort is unclear. Although there are no significant technical or biological differences between the BCCA and other sub-cohorts, it is worth considering the results from Figure 14 and Figure 20. Figure 14 indicates that the BCCA sub-cohort had fewer duplications than other sub-cohorts, while Figure 20 shows that chromothripsis was the only cSV type to have an increased risk in samples with more CNVs. Therefore, it is possible that the depletion of CNVs in the BCCA sub-cohort could explain the lack of chromothripsis depletion.

Other cSV types did not show significant enrichments or depletions in sub-cohorts, suggesting broadly similar distributions. Although tyfonas and seismic amplification were identified in the combined cohort, they were omitted from this analysis as they were present in fewer than 20 samples.

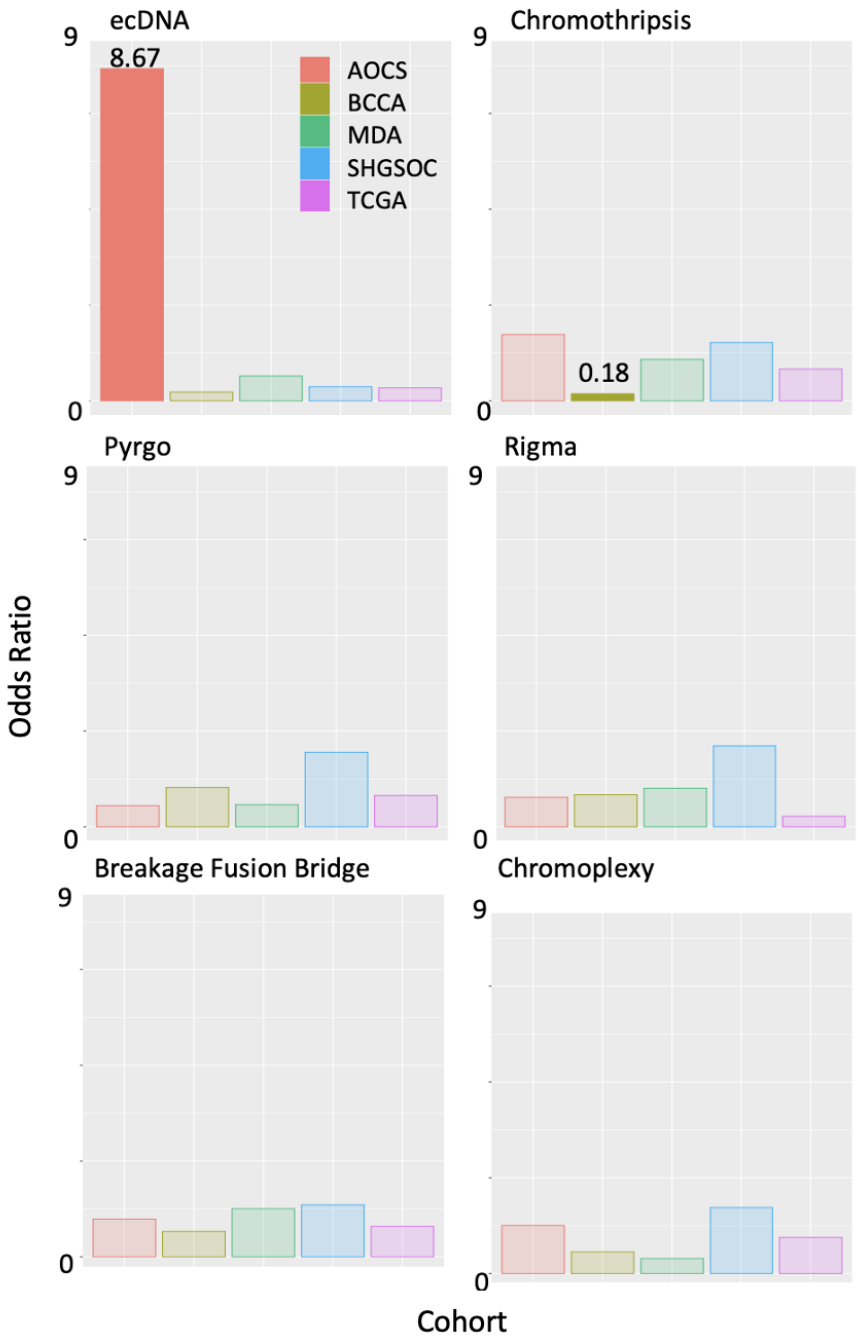


Figure 21 EcDNA is Enriched in the AOCS sub-cohort

For statistically significant results, the odds ratio is shown above the solid columns. The AOCS cohort was significantly enriched for ecDNA (Chi-squared test $p=4.59 \times 10^{-9}$,odds ratio of 8.67). Chromothripsis was found to be signif icantly depleted (Chi-squared test $p-4.1 \times 10^{-4}$, odds ratio of 0.18). Cohorts generating p-values greater than 0.05 are made translucent. Tyfonas and seismic amplification have been excluded due to being present in fewer than 20 samples.

Signatures of Underling Mutational Processes

The mutational processes that underly most cSV are poorly understood, but a recently developed algorithm attempts to identify the underlying mutational process (UMP) signatures associated with cSV (Bao et al. 2022; Carvalho and Lupski 2016). The Starfish algorithm infers UMP signatures and associates them with known genomic features; micronuclei, ecDNA and chromatin bridges (Bao et al. 2022). In addition, the Starfish algorithm infers three other mutational signatures that have not yet been linked to known mutational processes: large loss, large gain and hourglass chromothripsis (Bao et al. 2022).

Figure 22 shows the associations between cSV types and the inferred UMP signatures. The strongest co-occurrences identified were between the BFB (breakage fusion bridge) cSV type and the UMP signature associated with ecDNA, which co-occurred more than twice as often as expected by chance. This association was stronger than the relationship between the occurrence of the ecDNA cSV type and the UMP signature for ecDNA. The association between BFB and the UMP signatures for ecDNA may suggest a common mutational process driving the generation of ecDNA and BFB.

In addition, Figure 22 also confirms previously suggested associations between cSV types and UMP signatures. For example, it found a weak co-occurrence between chromothripsis and the UMP signature associated with micronuclei (C. Z. Zhang et al. 2015). Furthermore, Figure 22 b demonstrates that the UMP signatures detected in the combined cohort are distinct entities. The mutual exclusivity observed among these signatures indicates that they are separate and unrelated, consistent with findings in unrelated tumour samples (Bao et al. 2022).

Although Starfish represents an important step in linking cSV to UMP, it is not without limitations. The metrics used to determine UMP signatures were a modified version of the metrics used by ShatterSeek (Bao et al. 2022). These metrics were then used to describe all the samples within PCAWG. The optimal number of six clusters was then

determined based on multiple clustering algorithms. A single-layer neural net was then used to group samples into the six clusters (Bao et al. 2022). The UMP was then assigned by comparing the cSV identified in these samples by other cSV classifiers (Bao et al. 2022).

For example, in the UMP signature termed DM and ecDNA, 64 samples had ecDNA identified by AmpliconArchitect (Bao et al. 2022). However, there were 309 samples classified as the DM and ecDNA UMP signature that did not have ecDNA identified. Additionally, 36 samples where AmpliconArchitect identified ecDNA were not classified as the DM and ecDNA UMP signature (Bao et al. 2022). This raises questions as to the specificity and sensitivity of Starfish.

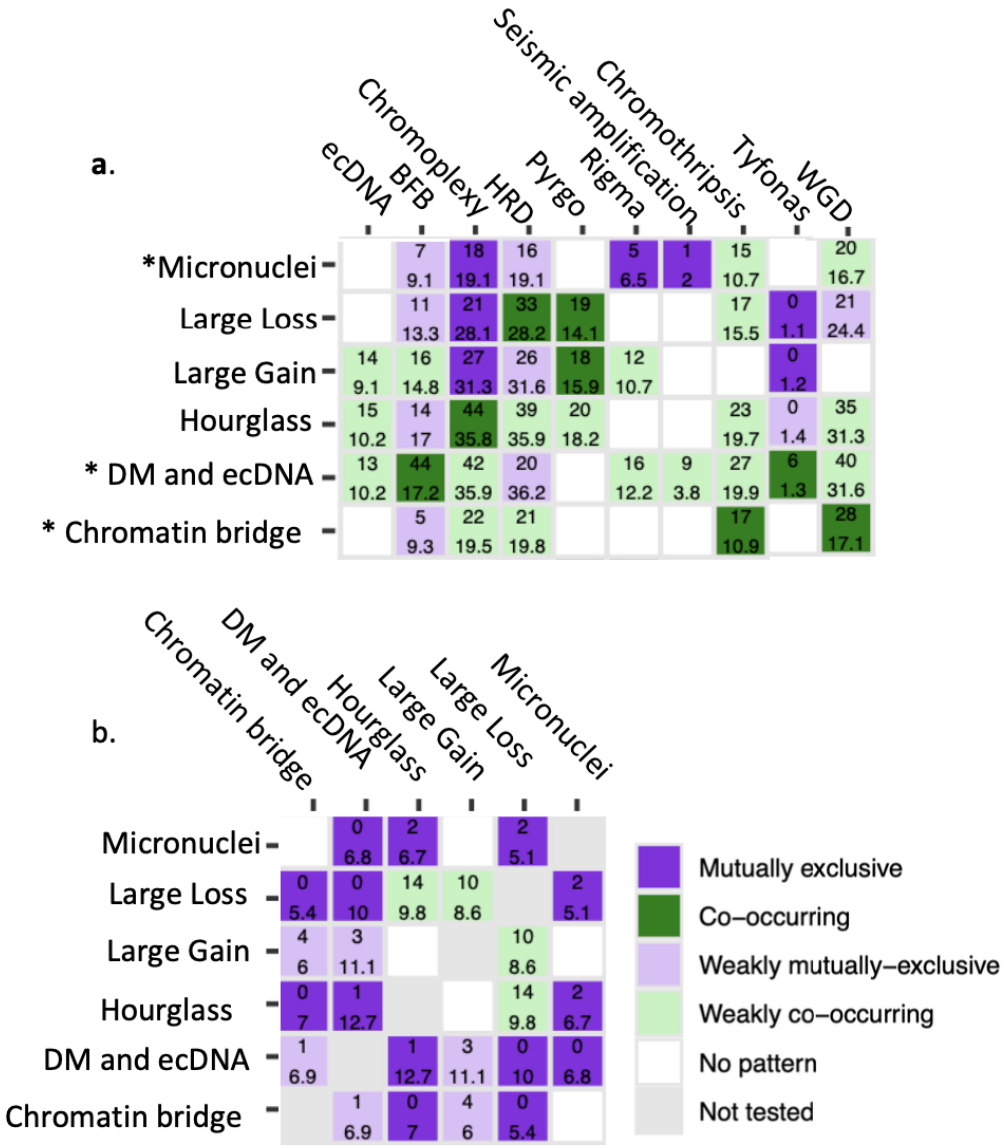


Figure 22 Inferring underlying mutational processes of complex structural variants

The observed (top number) and expected (bottom number) co-occurrence for each pair of cSV and mutational signatures. The complex structural variant identified in this work are the columns and the mutational signature and the rows. The rows with * are the underlying mutational process signature that are proposed to be biologically meaningful by the authors of Starfish (Bao et al. 2022; Carvalho and Lupski 2016) **b.** The pairwise co-occurrence of the mutational signatures within samples with the observed co-occurrence shown as the top number and expected co-occurrence shown as the bottom number.

The co-occurrence of complex structural variants suggest multiple evolutionary routes to structural diversity

Multiple cSV types are often identified in a single sample, and the co-occurrence of chromothripsis and ecDNA has been particularly widely reported (Shoshani et al. 2020; Robert and Crasta 2022; X.-K. Zhao et al. 2021a). Figure 23 shows the patterns of co-occurrence and mutual exclusivity between cSV types in the combined cohort, as well as the incidence of HRD and WGD. Overall, the occurrence of multiple cSV together in a sample was common, with only 13 (4%) samples having no detectable cSV and 175 (54%) samples having more than one cSV type. Figure 23 also suggests that certain cSV types may co-occur in samples more than expected by chance, which might result in clusters of samples sharing the same spectrum of cSVs (X.-K. Zhao et al. 2021b, 2021a; Shoshani et al. 2020)

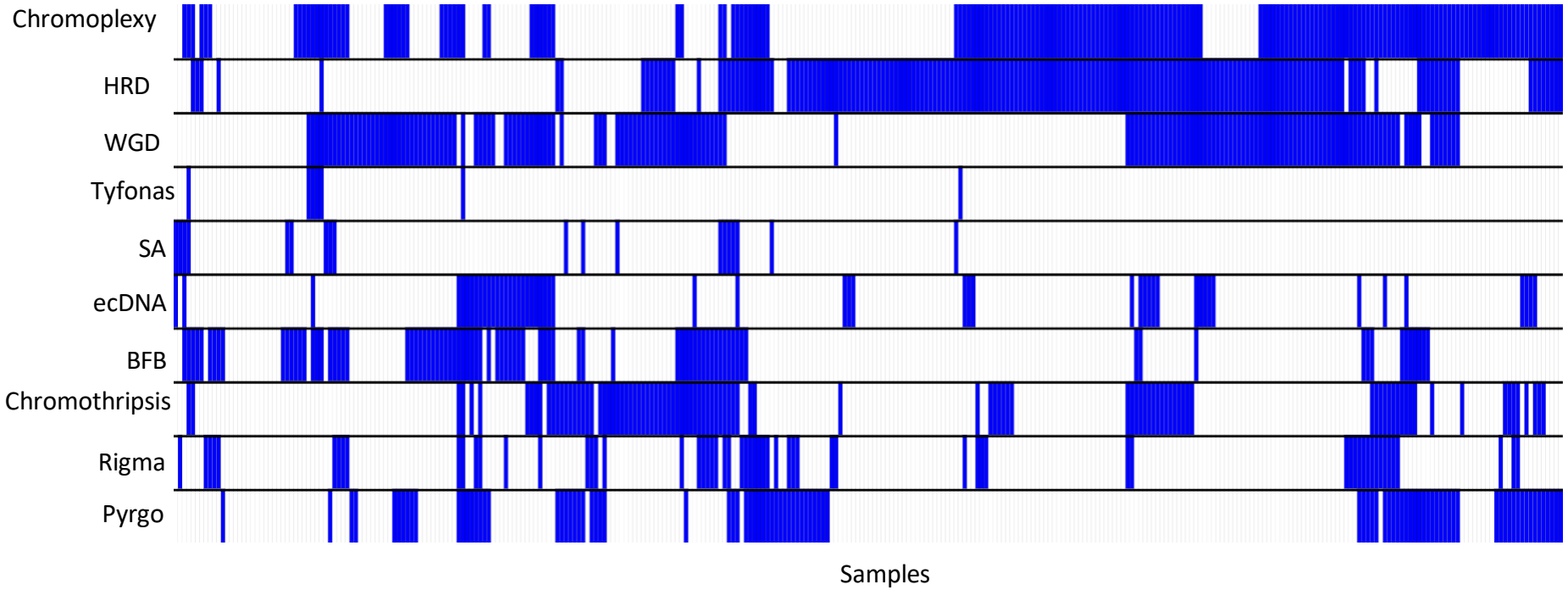


Figure 23 Overview of complex structural variant occurrence across the combined cohort

Each column represents a sample and the presence of each cSV in the sample is shown in blue. Breakage fusion Bridges (BFB), Homologous recombination repair deficiency (HRD) Whole genome duplication (WGD) and extrachromosomal circular DNA (ecDNA), seismic amplification (SA).

To investigate if there are clusters of complex structural variants that frequently occur together, hierarchical binary clustering was used with multiscale bootstrap resampling (Figure 24). The genomic complexity of samples was simplified to the present or absence of each of the CSV investigated. Although none of the clusters reached the recommended level of 95% support, two general groups appear to emerge. Chromoplexy and HRD cluster together with 86% support and the other cSV as well as WGD cluster together with 90% support, suggesting two broad evolutionary trajectories to generate structural diversity in HGSOC.

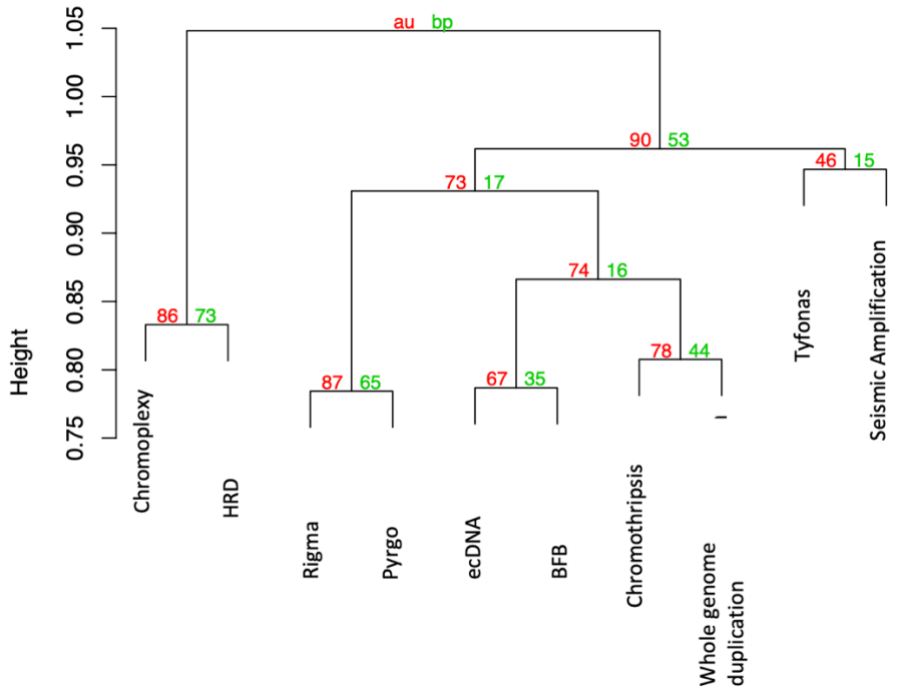


Figure 24 Sample clustering based upon cSV occurrence

The hierarchal clustering was done for 10000 permutations (shown on **left** in **Red)** has shown the approximately unbiased p-value calculated from multiscale bootstrap resampling. This test is reported to be superior to the commonly utilized bootstrap probability (shown on the **right** in **green)**. Cluster support was assessed by multiscale bootstrap resampling (Pvclust) based upon 10000 permutations (Suzuki and Shimodaira 2006).

Combination of Genomic Features	Number of samples
Chromoplexy, HRD	26
HRD	23
Whole Genome Duplication, chromoplexy, HRD	20
None	13
Whole Genome Duplication, HRD	10
Chromothripsis, Whole Genome Duplication, chromoplexy, HRD	8
Chromothripsis, Whole Genome Duplication, HRD	8
Pyrgo, HRD	7
chromoplexy	7
Chromothripsis, chromoplexy, HRD	6
Whole Genome Duplication, pyrgo, chromoplexy, HRD	6
Whole Genome Duplication	6
Whole Genome Duplication, BFB	5
Chromothripsis, Whole Genome Duplication	5
ecDNA, Whole Genome Duplication, BFB	4
Whole Genome Duplication, chromoplexy, BFB	4
ecDNA, Chromothripsis, Whole Genome Duplication, chromoplexy, HRD	4
Pyrgo, chromoplexy, HRD	4
Chromothripsis, Whole Genome Duplication, rigma, BFB	4

Table 4 Recurrent combinations of cSV appear in many samples

The table shows all the combinations of cSV that occur in more than 3 samples in the combined cohort.

Although the co-occurrence of cSV, and features such as HRD and WGD is evident across the combined cohort (Figure 23 and Figure 24), Table 4 shows that certain combinations are seen much more frequently. To directly investigate the co-occurrence or mutual exclusivity of cSV and other genomic features, the Select algorithm was used to compare the observed co-occurrence to a simulated null based on the proportion of observed samples with the cSV as shown in Figure 25 (Mina et al. 2017). Strikingly, this analysis suggests that HRD occurs exclusively from all cSV except chromoplexy which is weakly co-occurring but does not reach statistical significance, though often the magnitude of the relationships identified by Select appear to be modest. The strongest

exclusivity was seen for BFB, which was observed to co-occur with HRD in less than half the number of samples that was expected.

Most cSV types show some degree of co-occurrence with WGD except for prygo and seismic amplifications, which are modestly under-represented (Figure 25). BFB and ecDNA are shown to co-occur in the same sample twice as often as the Select algorithm's expectation**.** BFBs and ecDNA have similar defining features of amplified copy number and fold-back inversions. It is therefore possible that the observed co-occurrence of ecDNA and BFB is due to callers struggling to distinguish between the two types of events. Alternatively, they could be co-occurring events as proposed in the literature for ecDNA and chromothripsis (Shoshani et al. 2020).

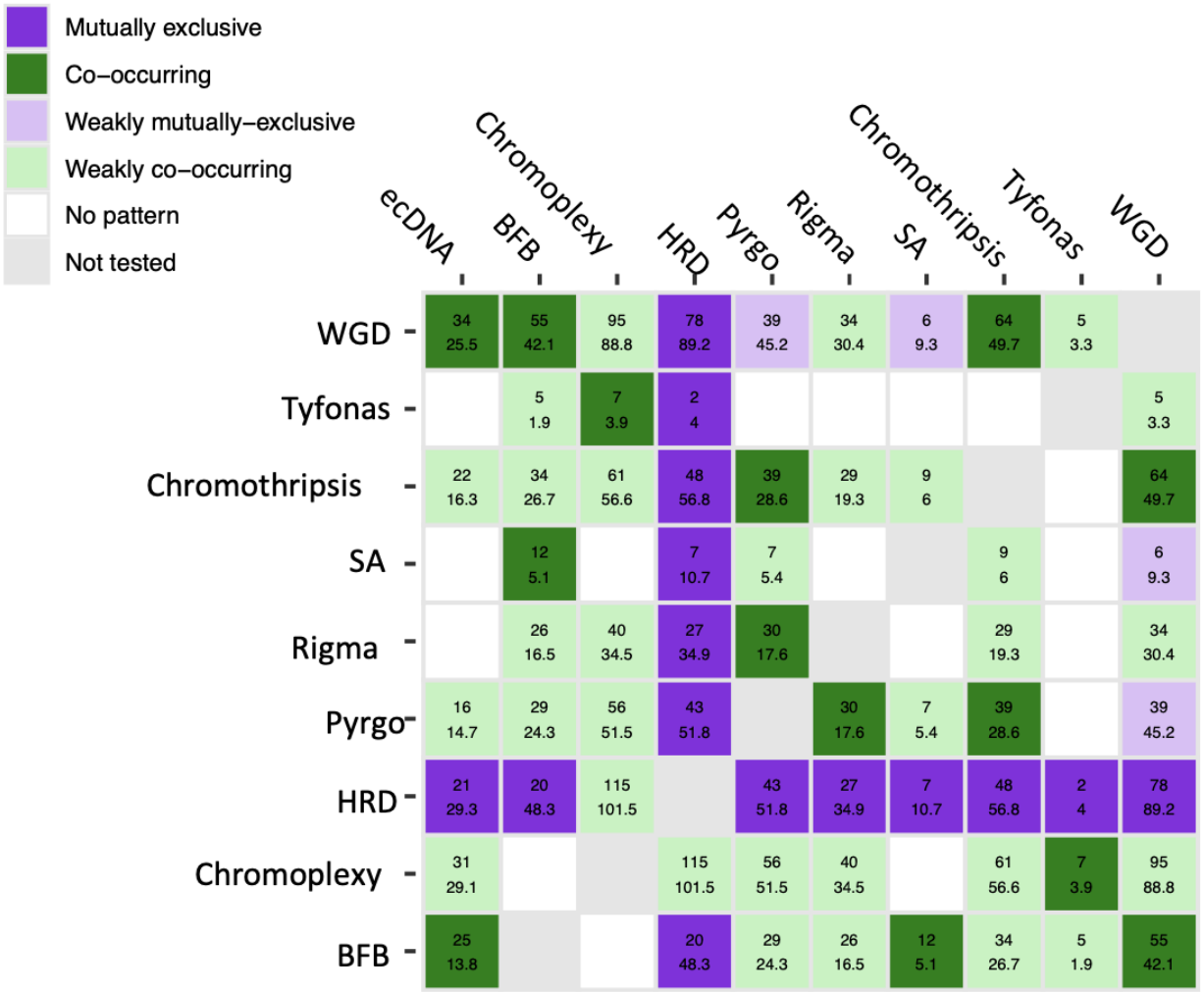
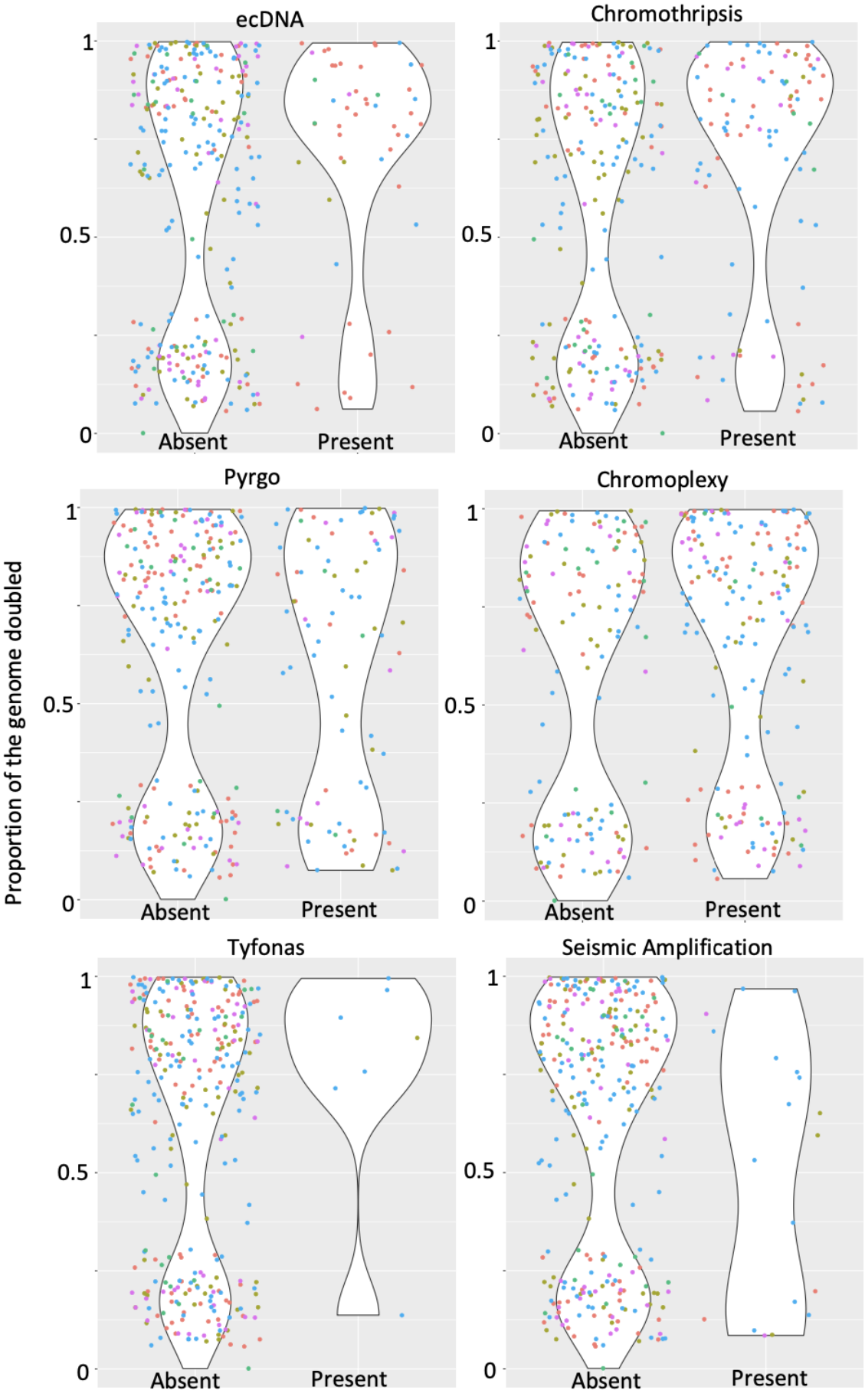
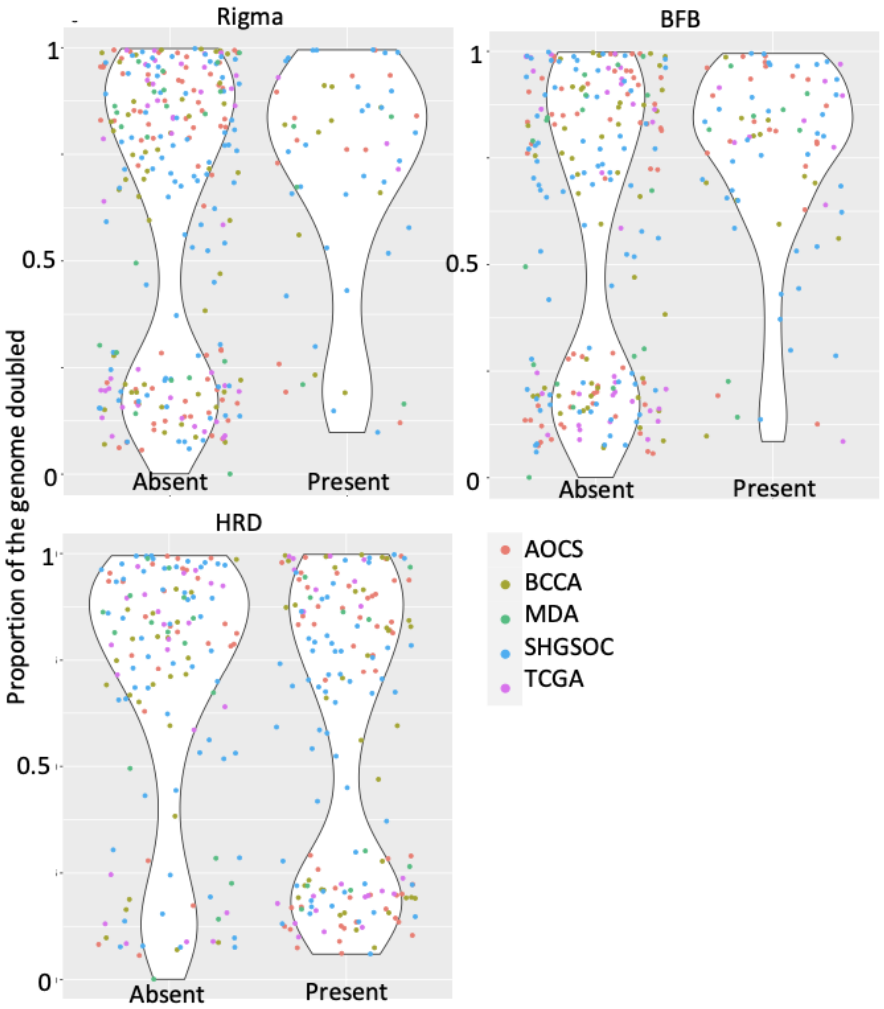


Figure 25 Trends in cSV binary co-occurrence and mutual exclusivity

The number at the top represents the observed overlap between the two cSV. The number below represents the predicted overlap between the two groups based on their frequency within samples.

To further investigate the relationship between WGD and cSV, we explored the quantitative data underlying the binary associations displayed in Figure 25, to examine the relationship between the proportion of the genome doubled in a sample and the occurrence of cSVs (Figure 26). These data suggest that samples with WGD are more than twice as likely to have chromothripsis, ecDNA and BFB than would be expected by chance (Figure 26). Additionally, HRD occurred in less than half of the samples with WGD than would be expected by chance. One interpretation of this is that WGD is an early event which allows cells to subsequently tolerate complex disruptive rearrangements such as chromothripsis and survive (Boisselier et al. 2018).


ecDNA
Chromothripsis
Pyrgo
Chromoplexy
Tyfonas
Seismic Amplification
Proportion of the genome doubled
1
0.5
0
Absent
Present



Complex structural variant	Pvalue	Odds ratio	UCI	LCI
ecDNA	0.13	2.22	4.39	1.15
Chromothripsis	$6.68x10^{-3}$	2.32	3.90	1.40
RIgma	1	1.33	2.42	0.74
Pyrgo	1	0.69	1.15	0.41
Chromoplexy	1	1.36	2.17	0.86
Breakage Fusion Bridge	0.014	2.28	3.95	1.34
Tyfonas	1	2.64	28.1	0.42
Seismic Amplification	1	0.46	1.34	0.14
Homologous recombination repair deficiency	0.12	0.57	0.90	0.35

Figure 26 The extent of WGD and cSV occurrence

The nine panels show the proportion of the genome that is doubled (y-axis) for samples

with and without each cSV type and HRD (x-axis); points represent samples coloured by sub-cohort of origin. The table shows the association between each cSV type and the extent of WGD assessed using Chi-squared tests with the Bonferroni correction. WGD was defined as samples with greater than 75% of the genome with a ploidy of 4.

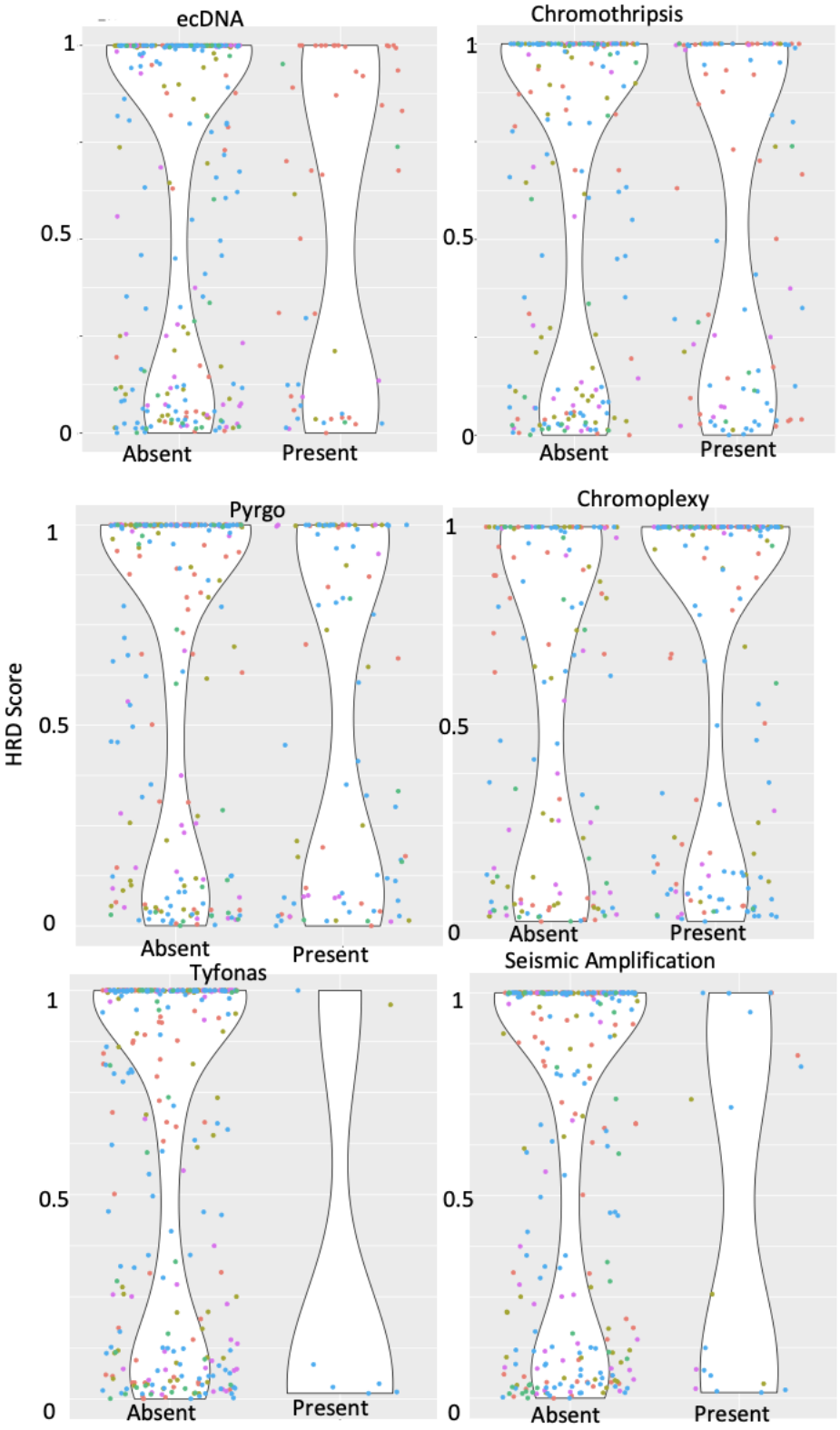
Overall the general trend was that cSV were significantly enriched in sample with WGD occurring more than twice as often as expected by chance (Table 5). This enrichment is driven mostly by chromothripsis. And breakage fusion bridge (Figure 26)

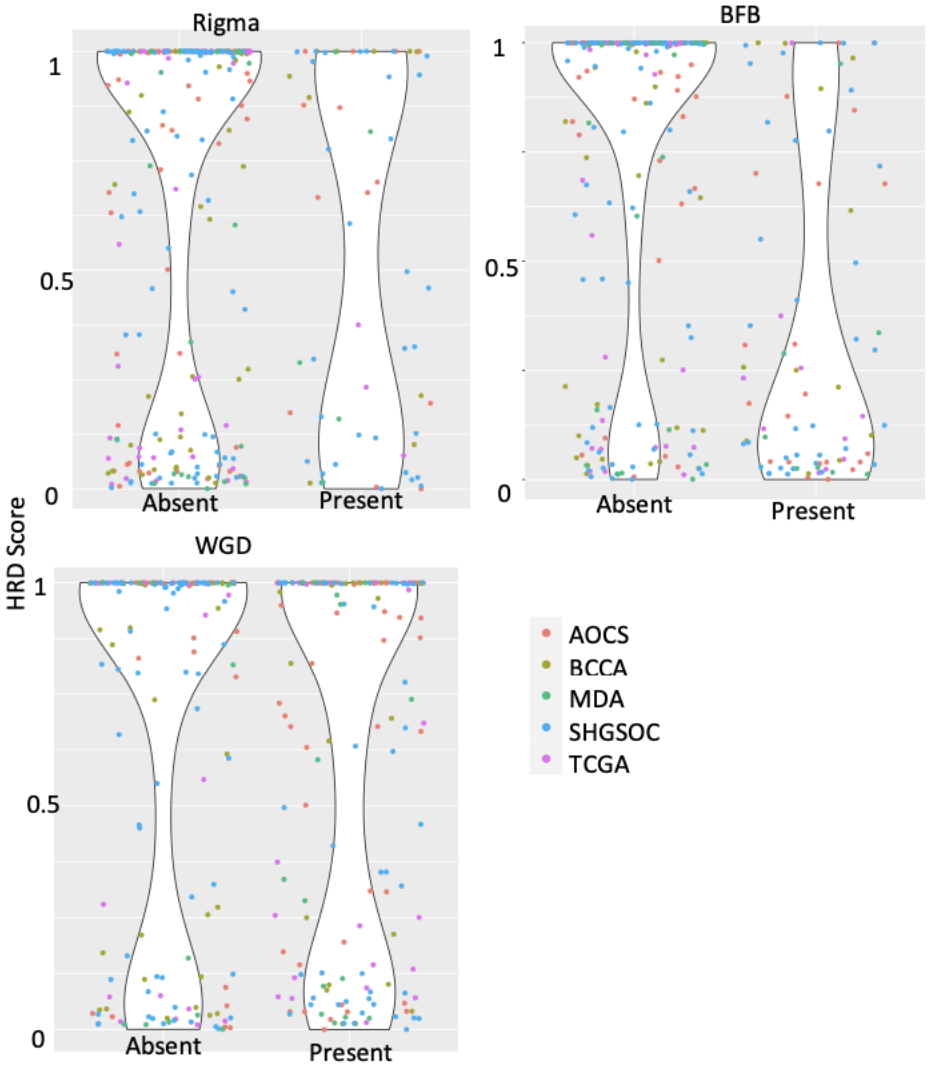
	Samples with WGD	Samples without WGD
Samples with cSV	142	129
Samples without cSV	17	36

Table 5 Enrichment of complex structural variants in samples with whole genome duplication

Number of samples with and without complex structural variants. (cSV) and whole genome doubling (WGD); Fisher's exact test p = 0.007, Lower confidence interval CI = 1.20, Upper confidence interval = 4.64, odds ratio = 2.33

The relationship between the extent of HRD and the occurrence of each cSV type was investigated analogously to WGD, exploiting the quantitative assessments of HRD (Figure 27). The most significant result was the depletion of BFB in HRD, such that only about 14% of the number of samples that would be expected by chance were observed. This depletion of BFB in samples with HRD has not been previously reported in the literature, though both HRD and BFB are known to increase genomic instability in tumours (Drews et al. 2022).


ecDNA
Chromothripsis
Pyrgo
Chromoplexy
Tyfonas
Seismic Amplification
HRD Score
1
0.5
0
Absent
Present



Complex Structural Variant	Pvalue	Odds ratio	UCI	LCI
ecDNA	0.13	0.468	0.89	0.242
Chromothripsis	0.36	0.603	0.993	0.364
RIgma	0.29	0.534	0.967	0.292
Pyrgo	0.32	0.588	0.984	0.35
Chromoplexy	0.030	1.97	3.17	1.23
Breakage Fusion Bridge	6.08×10^{-12}	0.143	0.259	0.0764
Tyfonas	1	0.306	1.9	0.0287
Seismic Amplification	0.87	0.434	1.24	0.141
Whole genome duplication	0.12	0.566	0.901	0.354

Figure 27 The extent of HRD and cSV occurrence

The HRD score for samples (y-axis) is related to the occurrence of each cSV type (x-axis) in the 9 panel Figure; points represent samples coloured by sub-cohort of origin. HRD scores based on the mutational signature of homologous recombination repair deficiency were estimated using the HRDetect algorithm (Davies et al. 2017). The table shows the association between each cSV type and the extent of HRD assessed using Chi-squared tests with the Bonferroni correction.

Chromoplexy was also found to be significantly (p=0.03) enriched in samples with HRD, occurring almost twice as often as would be expected by chance. A similar result of enrichment between chromoplexy and HRD has been previously shown in the literature in multiple myeloma (Ashby et al. 2018). This work defined chromoplexy as rearrangements within 1MB that involve at least 3 chromosomes from long read nanopore sequencing data and HRD was inferred from mutation in particular genes (TP53 and ATM) (Ashby et al. 2018). The same study also states that chromothripsis defined from manual curation is enriched in HRD. In contrast, in the HGSOC combined cohort studied here finds that chromothripsis is somewhat depleted, which may reflect real biological differences between cancer types or the different calling criteria employed in the current work and the previous study (Ashby et al. 2018).

To further investigate the general trend between cSV occurrence and HRD, the proportion of samples with any cSV (excluding chromoplexy and WGD) was compared to samples with HRD (Table 6), demonstrating that cSVs are three time more likely in samples without HRD. This is again consistent with HRD and cSV providing divergent routes to structurally diverse tumour genomes. However these two routes are not completely distinct, with almost half of HRD samples possessing some cSV, rather they seem to reflect general tendencies in tumour evolution. Additionally, samples with HRD were significantly depleted for WGD with roughly half the number of samples expected by chance (Table 7). It is also notable that a minority of samples did not possess HRD or cSV, highlighting that our knowledge of the origins of structural diversity is incomplete.

	Samples with cSV	Samples without cSV
Samples without HRD	80	29
Samples with HRD	102	113

Table 6 Complex structural variant enrichment in samples without homologous recombination deficiency

Samples with cSVs of any type excluding chromoplexy are three time more likely in samples without HRD; Fisher's exact test p = 1.01×10^{-05}, LCI = 1.80, UCI = 5.24, odds ratio = 3.05.

	Samples with HRD	Samples without HRD
Samples with WGD	81	78
Samples without WGD	106	56

Table 7 Whole genome duplication is depleted in samples with homologous recombination deficiency

Samples with cSVs of any type excluding chromoplexy are three time more likely in samples without HRD; Fisher's exact test p = 0.018, LCI = 0.36, UCI = 0.92, odds ratio =0.58.

The role of complex structural variants in generating genomic complexity

In Figure 16 – Figure 18 it was shown that cSV can encompass multi-mega base regions and may contain hundreds of CNVs and simple SVs. This raises the possibility that cSV may play a major role in generating the complex rearranged HGSOC genome. More specifically, we can ask what fraction of the total SVs and CNVs seen within HGSOC samples can be explained by the known cSV types examined above. Since our knowledge of cSV is likely to be incomplete, we can also estimate the SV/CNV fraction that could conceivably be explained by all known and unknown cSV types. As a proxy for all cSV we use significant clusters of SV/CNV detected using established methods, since all known cSV types involve clustered SVs (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (89ormalized89n/Ins et al. 2020).

Proportion of Structural Variants explained by known cSV types

To investigate what proportion of SV across the combined cohort could be explained by the presence of known cSV, the number of consensus SV within regions of a cSV were divided by the total number of SV identified across the combined cohort (Table 8). By this measure the cSV explain 13,877 consensus SV calls, or 14.56% of all SV, representing a small fraction of the total structural diversity present within the combined cohort. The proportions of SV calls of each type explained by cSV are as expected based on the criteria for calling each cSV type. For example, pyrgo (which is enriched for duplications) explains more SV duplications than deletions and inversely rigma (which is enriched for deletions) explains more deletions than duplications. Similarly, chromoplexy is defined as chains of reciprocal translations and, as expected, explains the largest proportion of translocations of any cSV type.

Complex Structural Variants	Any SVs Explained	Deletions Explained	Duplications Explained	Translocations Explained	Inversions Explained
All cSV	13877 (14.56%)	3457 (8.87%)	2562 (10.10%)	4575 (26.78%)	3279 (23.66%)
Chromothripsis	5972 (6.27%)	1690 (4.34%)	1480 (5.83%)	887 (5.19%)	1913 (13.80%)
BFB	2956 (3.10%)	554 (1.42%)	424 (1.67%)	886 (5.19%)	1091 (7.87%)
Chromoplexy	3648 (3.83%)	844 (2.17%)	0 (0%)	2804 (16.42%)	0 (0%)
ecDNA	1225 (1.29%)	205 (0.53%)	262 (1.03%)	338 (1.98%)	420 (3.03%)
Tyfonas	850 (0.89%)	141 (0.36%)	103 (0.41%)	391 (2.29%)	214 (1.54%)
Pyrgo	618 (0.65%)	35 (0.09%)	524 (2.06%)	23 (0.13%)	36 (0.26%)
Rigma	578 (0.61%)	526 (1.35%)	23 (0.09%)	7 (0.04%)	22 (0.16%)
Seismic Amplification	431 (0.45%)	116 (0.30%)	110 (0.43%)	31 (0.18%)	174 (1.26%)

Table 8 Proportion of structural variants explained by complex structural variants

The number and proportion of structural variants across the combined cohort explained by each known cSV type.

Although cSV explains a modest fraction of SVs across the combined cohort, there is substantial variation in the proportion explained between different samples. The proportion per sample was calculated (Figure 28) and was found to vary widely from 0% in samples that do not contain a known cSV to 100% in one outlier sample. Overall, cSV explain more than 50% of SV calls in a small minority (1.85%) of samples, but the mean over all samples is 10.95% of SV calls explained by cSV. A similar result was found for CNV, where cSV explained an even smaller proportion of CNV (4.35% on average) in samples (Figure 29). We conclude that in spite of the often catastrophic rearrangements associated with cSV, they generally account for only a small fraction of the total structural diversity seen in any given HGSOC sample.

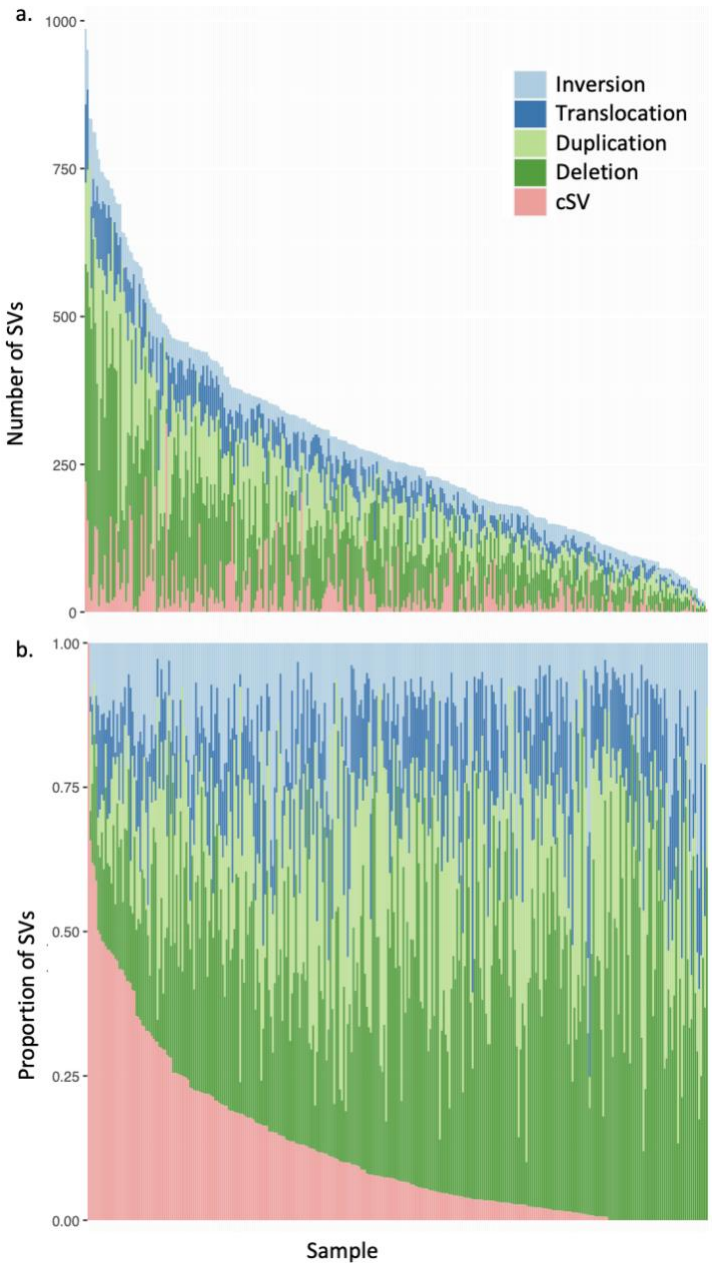


Figure 28 Complex structural variants explain a variable proportion of SV calls across samples

a. The number of consensus structural variants (SVs) explained by known complex structural variants (cSVs) for each sample (pink) is depicted with the remaining SVs (blue, green) in each sample that are not associated with cSV. The graph is ordered by total number of SVs **b.** The proportion of consensus SV in each sample that are explained by known cSVs (pink) with the remaining SVs (blue, green) ordered by proportion of SV explained by cSVs.

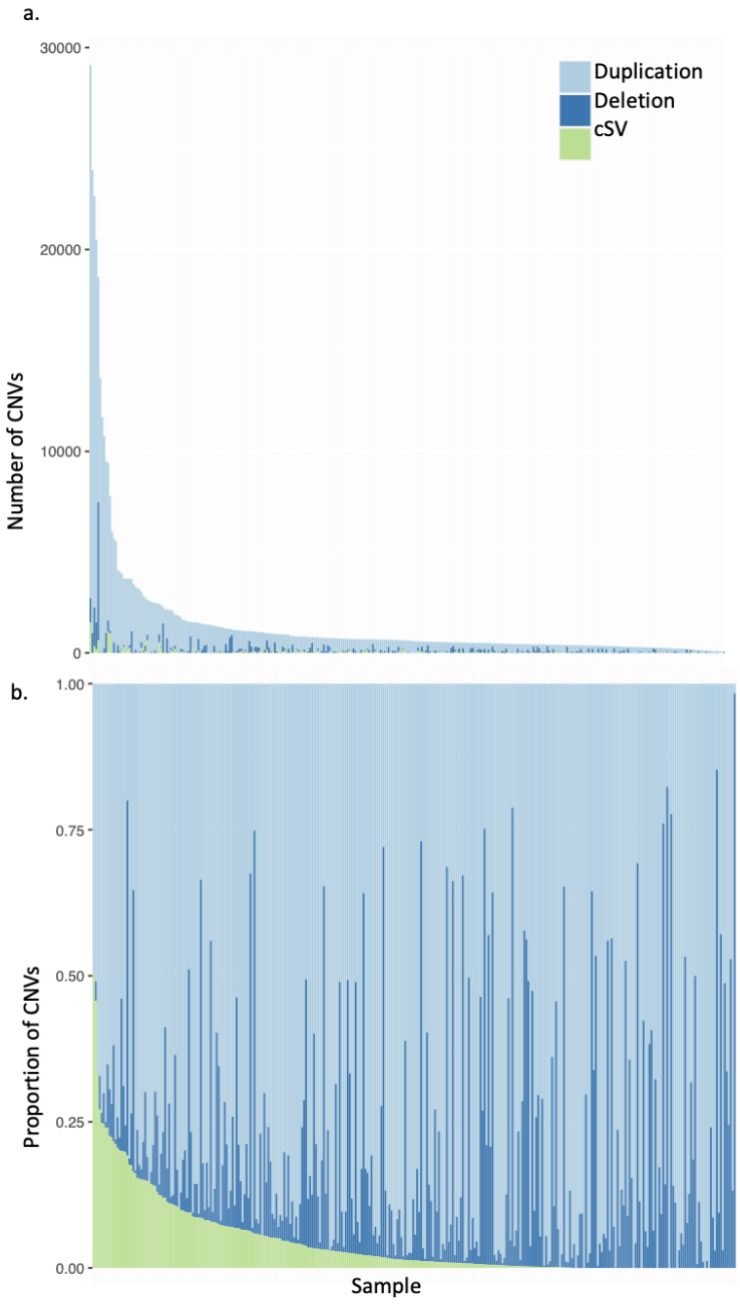


Figure 29 Complex structural variants explain a modest fraction of CNV calls across samples

a. The number of consensus copy number variants (CNV) explained by known complex structural variant (cSV) types (green) for each sample with remining CNV calls not associated with cSV (blue). Overall, cSV explained 15,619 CNV calls (3%) across the combined cohort, ordered by number of CNVs. **B.** The proportion of CNV in each sample that are explained by cSV (green) and CNV not associated with cSV (blue), ordered by the proportion of CNVs explained.

Currently unexplained clusters of SVs may represent novel cSV

As mentioned previously, our understanding of cSV is far from complete. Many cSV types were only discovered recently (Chapter 1), most cSV callers rely upon arbitrary thresholding of SV and CNV data. Furthermore, the ability to resolve cSV is likely to be limited by the technical constraints of the short read WGS data that exists for the vast majority of tumour samples. This raises the question: how much more complex structural variation might exist in tumours? One approach to address this is to identify significant clusters of SV calls that cannot be explained by any known cSV type, using existing cSV callers.

Significant clusters of SV calls were identified in the combined cohort using the same criteria as the ICGC PCAWG consortium which defined unexplained complex regions as those containing SV that were clustered more than expected (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (93ormalized93n/Ins et al. 2020). This resulted in 904 SV clusters in total, and each cluster was then classified as 'explained' (where a known cSV call overlapped more than 50% of the SV breakpoints within that cluster) or 'unexplained' (if not). Overall, 1,429 of the SV clusters (16%) were classified as explained known cSV types.

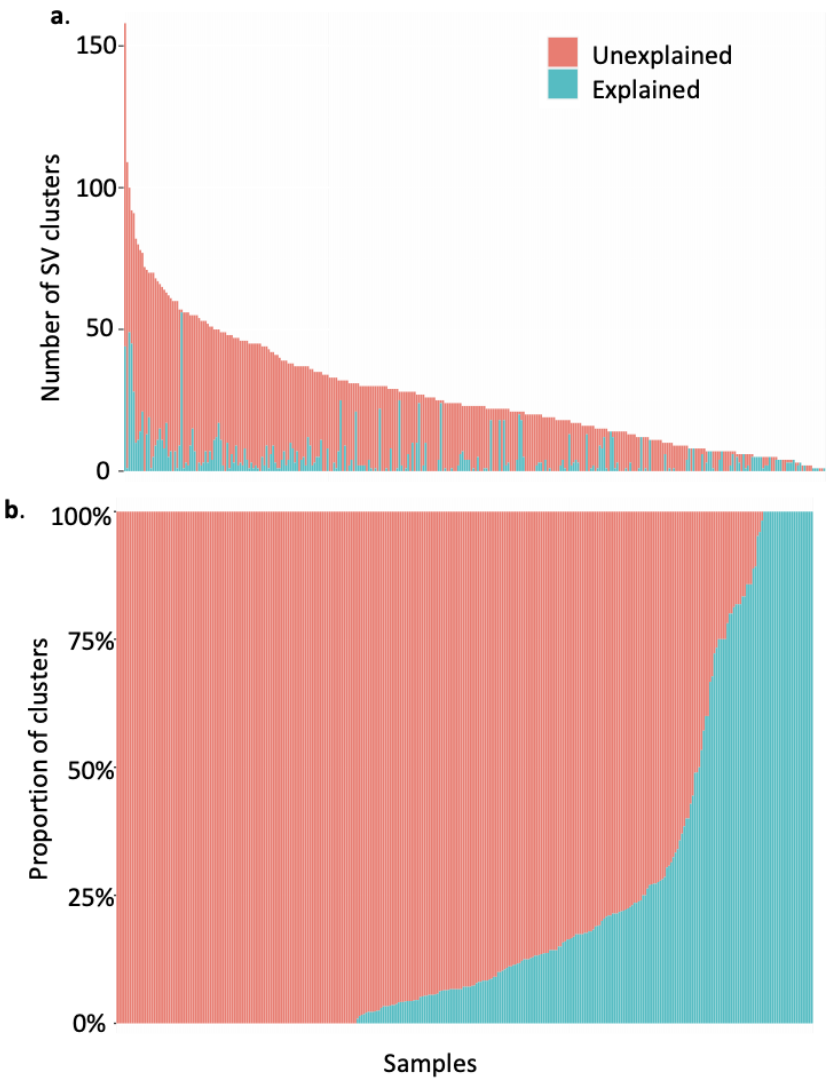
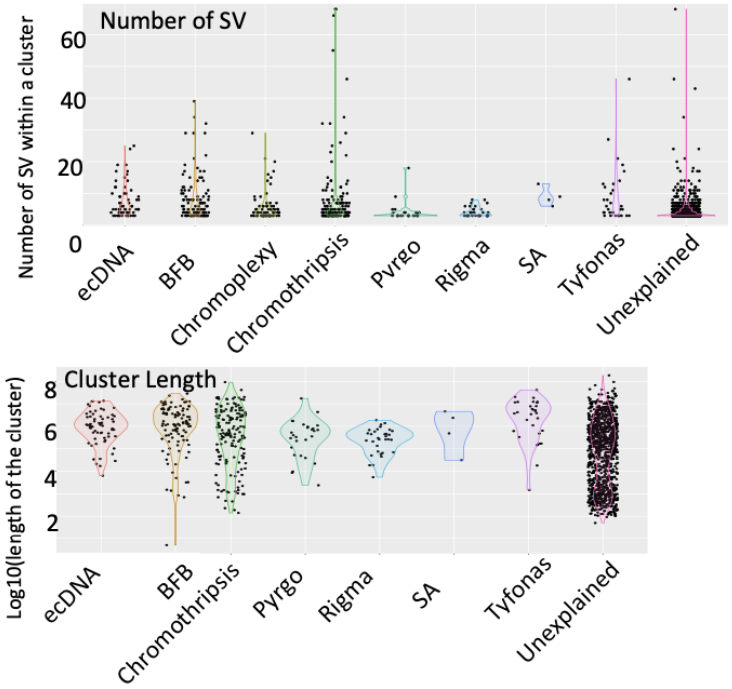


Figure 30 Known cSV types and unexplained SV clusters across samples

The number of structural variants (SV) clusters explained by known cSV types (blue) relative to the number that remain unexplained (red). An overlap of at least 50% between the SV breakpoints within each SV cluster was required to assign clusters to the 'explained' category (16% of clusters). Samples within the combined cohort are ordered by number of SV clusters. **B.** Proportion of SV clusters in a sample explained by known cSV types (blue) and those that remain unexplained (red). Ordered from least proportion of samples explained to greatest.

Figure 30 shows that cSV explain a majority of the structural complexity in very few samples across the cohort, suggesting that novel cSV events and perhaps novel cSV types may be present in these data, represented as unexplained SV clusters. These unexplained clusters are further described in terms of their numbers of SV and cluster size in Figure 31 and Figure 32, and compared to the clusters assigned to each known cSV type.



Number of SVs		
cSV	Median	IQR
Unexplained	3	2
BFB	5	6
ecDNA	5	6
Chromothripsis	5	5
Rigma	3.5	2
Chromoplexy	4	3
SA	9	13
Pyrgo	3	1
Tyfonas	6.5	9

Cluster Length (BP)		
cSV	Median	IQR
Unexplained	135764	1123437.5
BFB	1270229	326334
ecDNA	1120417	2371222
Chromothripsis	742029	3663655
Rigma	261036	485514
SA	3222499	5380080.5
Pyrgo	376691	833647
Tyfonas	2361569.5	5037662.5

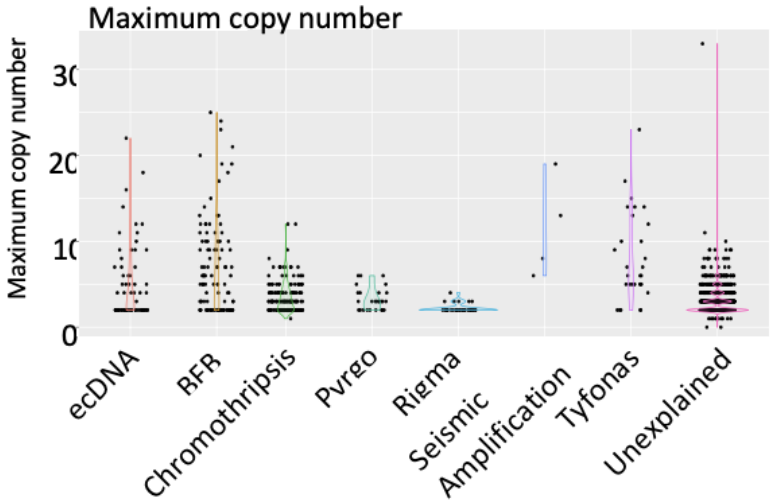
Figure 31 Clusters that are explained by cSV are not distinct from unexplained clusters Shows the length of cluster and number of SV involved in the clusters that are explained by different cSV types as well as the clusters that remain unexplained. With median and interquartile range for each measure of the cluster shown by type.

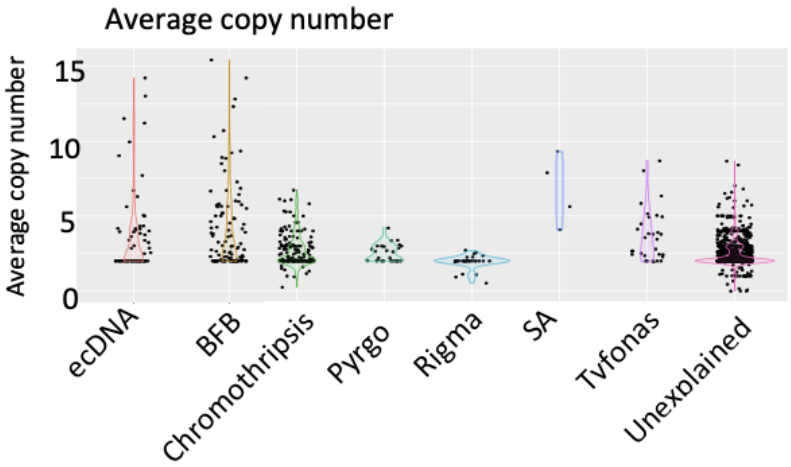
The number of SV and length of unexplained clusters is similar to the number of SV and length of clusters that are explained by cSV. Figure 31 shows that the majority of cluster of all classifications have relatively few structural variants. Some clusters were observed to have tens of SVs. The most simplistic definition of a complex structural variant is multiple SV clustered closer together than chance. However, the presence of unexplained clusters containing tens of SVs suggests that our current understanding of cSVs is incomplete.

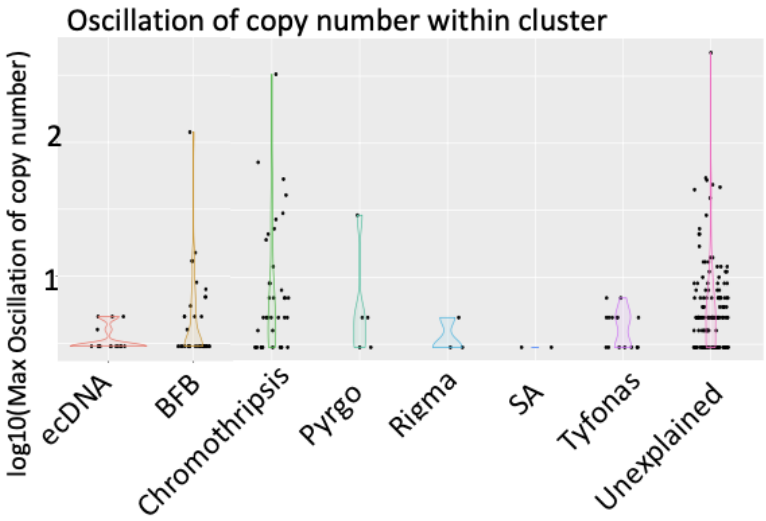
Clusters of SV can also be investigated by looking at the copy number within the clustered regions. Maximum and average copy number within these regions as well as any oscillation of copy number (a feature of complex mutational events involving many simultaneous rearrangements, such as chromothripsis) are shown in Figure 32**.**

The majority of clusters have normal average and maximum copy number as shown in Figure 32**.** However, it is also clear that some unexplained clusters have unusually high amplified maximum and average copy number. Depending on the region of the genome that is amplified, this increased copy number could have important biological consequences.

The majority of clusters do not show oscillation. Therefore, to investigate oscillation within clusters, the investigation was restricted to clusters that showed some oscillation. As would be expected, clusters explained by chromothripsis showed oscillation as oscillation of copy number which is a key characteristic of chromothripsis. However, other clusters, including unexplained clusters, also show evidence of oscillation. Oscillation of copy number was proposed by Korbel and Campbell to be unlikely to occur by sequential structural rearrangements and instead indicated the simultaneous generation of these variants clusters may have occurred in a single cell cycle (Korbel and Campbell 2013). The presence of unexplained clusters with tens of SV, amplified copy number and oscillation of copy number strongly suggests that the landscape of complexity is far from fully annotated.


Maximum copy number
Maximum copy number
30
20
10
0
ecDNA
BFB
Chromothripsis
Pyrgo
Rigma
Seismic Amplification
Tyfonas
Unexplained


Average copy number
Average copy number
15
10
5
0
ecDNA
BFB
Chromothripsis
Pyrgo
Rigma
SA
Tyfonas
Unexplained


Oscillation of copy number within cluster
log10(Max Oscillation of copy number)
2
1
ecDNA
BFB
Chromothripsis
Pyrgo
Rigma
SA
Tyfonas
Unexplained

Maximum copy number		
cSV	Median	IQR
Unexplained	2	2
BFB	5	8
ecDNA	3	5
Chromothripsis	3	2
Rigma	2	0
SA	8	7
Pyrgo	3	2
Tyfonas	5	8.5

Average Copy Number		
cSV	Median	IQR
Unexplained	2	0.92
BFB	2.76	3.18
ecDNA	2.49	2.15
Chromothripsis	2.27	1.21
Rigma	2	0
SA	5.61	2.62
Pyrgo	2.41	0.99
Tyfonas	2.58	2.58

Oscillation of copy number		
cSV	Median	IQR
Unexplained	5	4.5
BFB	3	2
ecDNA	3	0.5
Chromothripsis	5	5.5
Rigma	3	1
SA	3	0
Pyrgo	5	2
Tyfonas	5	2

Figure 32 Some unexplained cluster have features associated with single cell cycle catastrophic events

The vast majority of clusters do not show oscillation therefore the Median oscillation of copy number is 0 but in the sample with at least 1 oscillation of copy number so the values represent the clusters with oscillation greater than 0. Copy number calls used for this analysis where the CNVkit calls as the consensus where too conservative to be informative for this work.

Known cSV types may be linked by shared underlying SV calls

In investigating the cSV explanations of structural variants, it became clear that some SVs could be explained by more than one cSV event as the same region of the genome can be identified as multiple cSV which could be interpreted as one region of the genome undergoing multiple complex events. However, as the criteria for calls of cSV

are not mutually exclusive it is possible that a complex region generated from one event is being called as two cSV events.

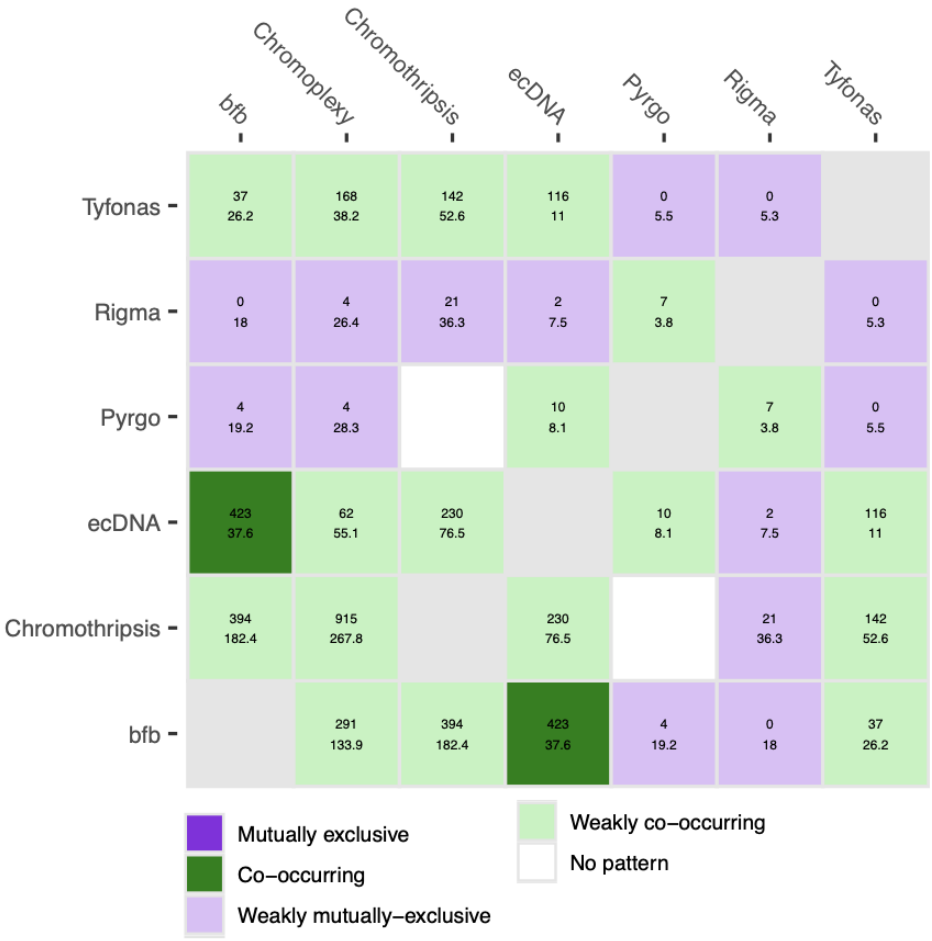


Figure 33 Shared underlying SV calls between different cSV types

In each cell in the interaction matrix, the number at the top represents the observed overlap of SV calls between the two cSV. The number below represents the predicted overlap of SV calls between the two groups based on their frequency within the population. Significance at the level of $p < 0.05$ (chi sq test) is indicated by the dark colours.

Figure 31 – Figure 33 all show that the current understanding and classification of complex structural event is not collectively exhaustive or mutually exclusive and although this will be further investigated in the coming chapters, it is an important

though unavoidable caveat.

Discussion

In this chapter, the prevalence and co-occurrence of eight cSV types has been explored across 324 whole genome sequenced HGSOC samples. By exploiting the well annotated nature of the combined cohort, this chapter was able to study the relationship between cSV and WGD and HRD.

The breadth of complex structural variants identified within the combined cohort has allowed for a deeper investigation into a general pattern of mutual exclusivity between cSV and HRD, for the first time hinting at two general pathways to genomic variation in HGSOC. The exception to this was the enrichment of chromoplexy in samples with HRD which is supported by the observation of increased translocations in HRD samples, probably due to the alternative end joining repair pathway used by HRD samples (Konstantinopoulos et al. 2015).

Although large numbers of clusters of SVs were identified in the PCAWG studies, the clusters were not investigated for the presence of key features of complexity (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (101ormalized101n/Ins et al. 2020). Here, I have shown that despite the breadth of cSV identified, the majority of complexity within the genomes of HGSOC remains unexplained. This unexplained complexity often exhibits typical characteristics of complex events, such as copy number oscillations (Korbel and Campbell 2013). This highlights how our current patchwork understanding of genomic complexity is far from comprehensive of complexity within the HGSOC. Additionally, the proportions of total SVs explained by cSV has not previously been reported in any cancer type for any cSV.

This chapter also demonstrated the enrichment of ecDNA within the AOCS sub-cohort which was selected for resistance to treatment. This sub-cohort was not significantly enriched for any other cSV or simple SVs. It was only through the combined analysis of the AOCS sub-cohort with other sub-cohorts that the enrichment of ecDNA in AOCS

could be determined. This highlights the value of the combined cohort approach, which allows for the emergence of new insights that may not have been possible otherwise.

Using SV numbers as a measure of genomic instability, it was found that there is generally no association between genomic instability and the presence of cSV calls. This suggests that samples with identified cSVs are not necessarily the most genomically unstable. Ultimately this chapter has shown the co-occurrence of cSVs across the combined cohort at a sample level is not random. In the next chapter, the occurrence cSVs across the genome will be investigated.

Chapter 4: Non-Random Patterns of Complex Structural Variants Across the Genome

Introduction

In the previous chapter complex structural variants (cSVs) were shown to have a non-random distribution across samples. In this chapter the distribution of cSVs across the genome will be investigated to detect unexpected hotspots of cSVs and to assess the impact on known high grade serous ovarian cancer (HGSOC) oncogenes and tumour suppressors.

The distribution of cSV across the genome is understudied. Some previous work has suggested that large chromosomes were more likely to be hit by chromothripsis. However, this work did not appropriately account for chromosome length (Klaasen et al. 2022). Here for the first time, I explore the enrichment of chromothripsis and other known cSV types accounting for chromosome length.

Utilising a previously published list of genes of clinical interest specifically in HGSOC (Hollis et al. 2022). As well as broader pan cancer lists of oncogenes, tumour suppressors and fusion genes from the Cancer Gene Census, the relationship between cSV and these genes of interest will be assessed (Sondka et al. 2018).

The key questions addressed in this chapter are:

i. Distribution: Are cSVs distributed uniformly across chromosomes, or are certain chromosomes enriched or depleted for cSVs?

ii. Hotspots: Are there specific regions in the genome where cSVs occur more frequently than expected by chance?

iii. Fragile Sites: Are cSVs more likely to occur in known fragile sites across chromosomes, and if so, which specific sites show the strongest association?

iv. Oncogenes: Is there an association between cSVs and oncogenes known to be relevant in ovarian cancer, and if so, which specific genes show the strongest association?

Simple Structural Variants are Unevenly Distributed across Chromosomes

Before investigating the distribution of cSVs across the genome, the distribution of structural variants (SVs) must first be assessed (Figure 34). The most enriched for total SV as well as each individual SV type was chromosome 19. Also enriched for total SV were chromosomes 8 followed by 20 and 12. Furthermore, chromosome 13 was the most depleted for total SV. There is a similar pattern of enrichment and depletion across the chromosomes for duplication and deletions. However, translocations were significantly depleted on every chromosome except chromosome 19 which was significantly enriched for translocations.

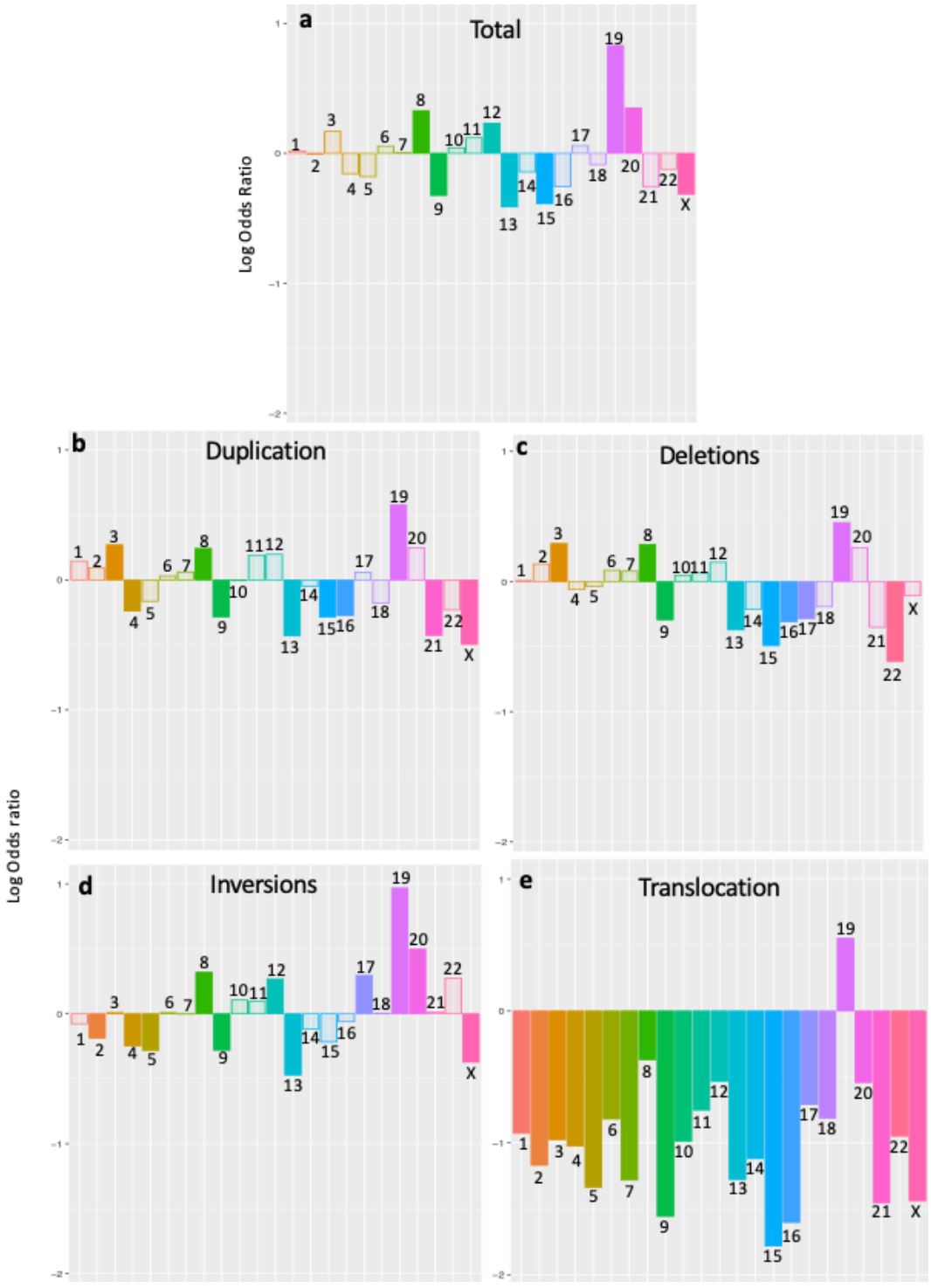


Figure 34 Distribution of SVs across chromosomes is not random

Observed occurrence of SVs across chromosomes relative to expectation, based on chromosome length, for (a) occurrence of total SVs, (b) duplications, (c), deletions, (d) inversions, I translocations. Positive values (y-axis) of log odds ratios indicate relative

enrichment and negative values indicate depletion. Significance was assessed using chi-squared tests ($p<0.05$ after Bonferroni correction) and is indicated by solid colours, translucent colours indicate a lack of significance.

The distribution of copy number variants (CNVs) across chromosomes is also non-random (Figure 35). Once again, chromosome 19 showed the most extreme effect size as it was the most depleted for total CNV SV and was also significantly depleted for CNVs of all types. However, the majority of chromosomes showed significant depletion (11 chromosomes) or enrichment (7 chromosomes).

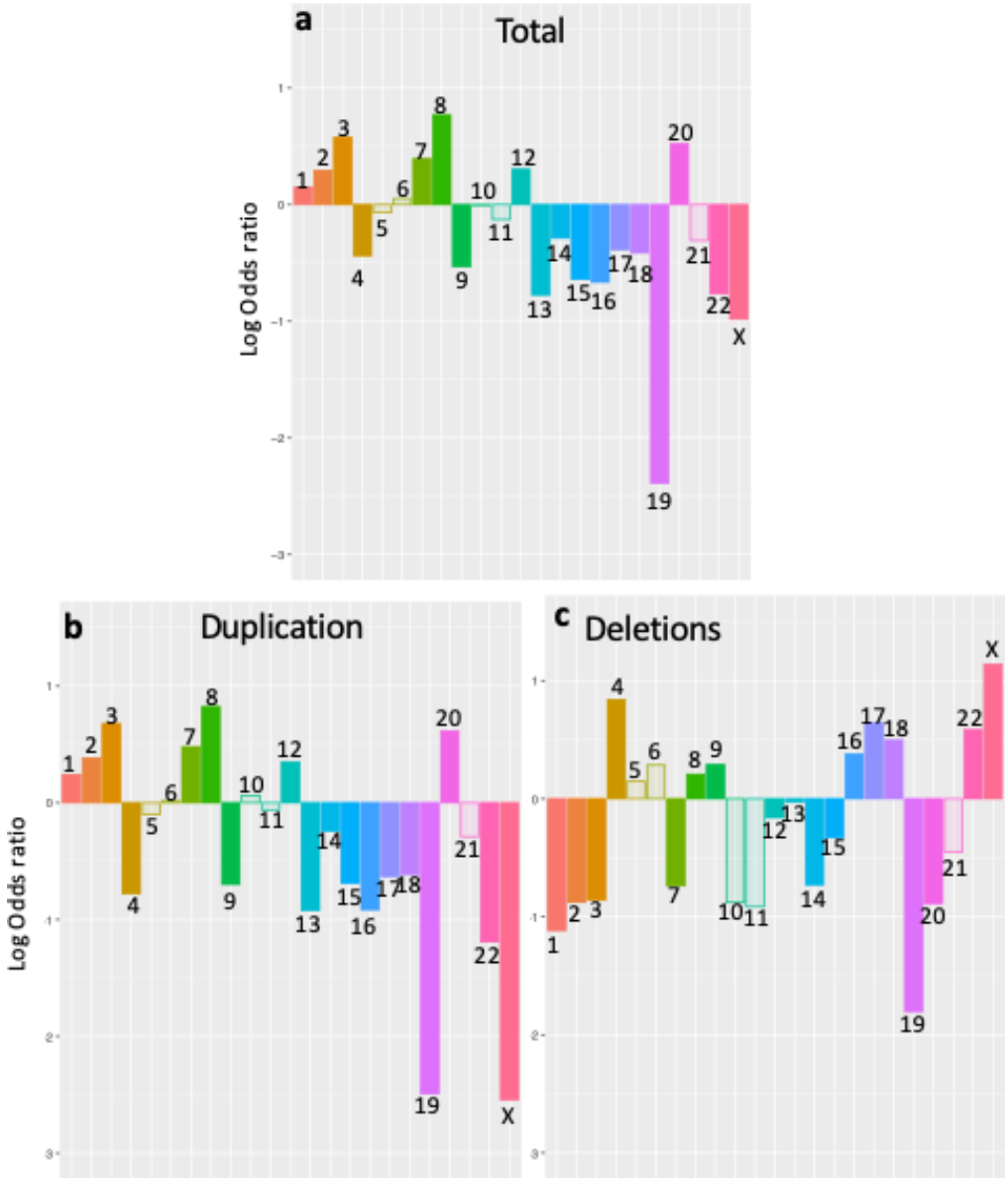


Figure 35 Distribution of CNV are not random

Observed occurrence of CNVs across chromosomes relative to expectation, based on chromosome length, for (a) occurrence of total CNVs, (b) duplications, (c), deletions. Positive values (y-axis) of log odds ratios indicate relative enrichment and negative values indicate depletion. Significance was assessed using chi-squared tests ($p<0.05$ after Bonferroni correction) and is indicated by solid colours, translucent colours indicate a lack of significance.

From Figure 34 and Figure 35 it is difficult to see if there is any relationship between the enrichment of SVs and CNVs across chromosomes. To investigate this, the enrichment of SVs and CNVs per chromosome was assessed (Figure 36). A strong positive correlation

between enrichment of total SVs and total CNVs was observed. However, chromosome 19 is an outlier showing the greatest enrichment of SV and the greatest depletion of CNVs.

Despite the strong positive correlation between total SV and CNV (Figure 36 b) shows a weaker but still significant negative correlation between enrichment of SV deletions and CNV deletions. The strong positive correlation of total SV and CNVs is driven by the strong positive correlation between SV and CNV duplications as shown Figure 36 c. The enrichment of duplication SVs and CNVs also showed that chromosome 19 was an outlier being enriched for SV and depleted for CNVs. Chromosome X was the most depleted for SV and CNV duplications. In the previous chapter it was shown that there are many more duplication than deletions (Figure 15). This difference in the quantity of deletions and duplications may explain why the total trend more closely follows the trend in duplications than deletions. Alternatively, it may be that SV callers encounter technical difficulties in calling deletions where larger CNV mediated deletions are also present, and these difficulties are less severe for duplications.

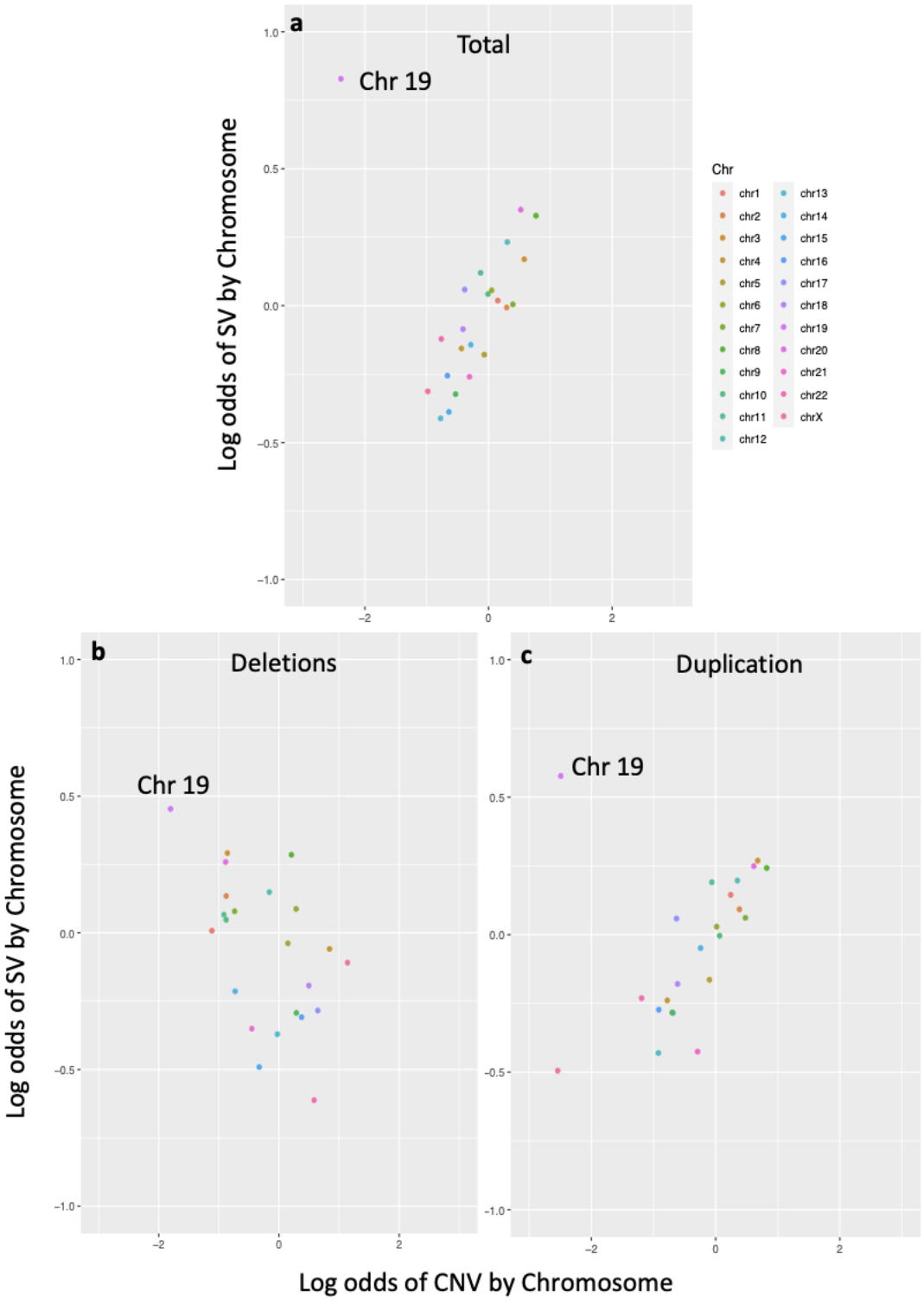
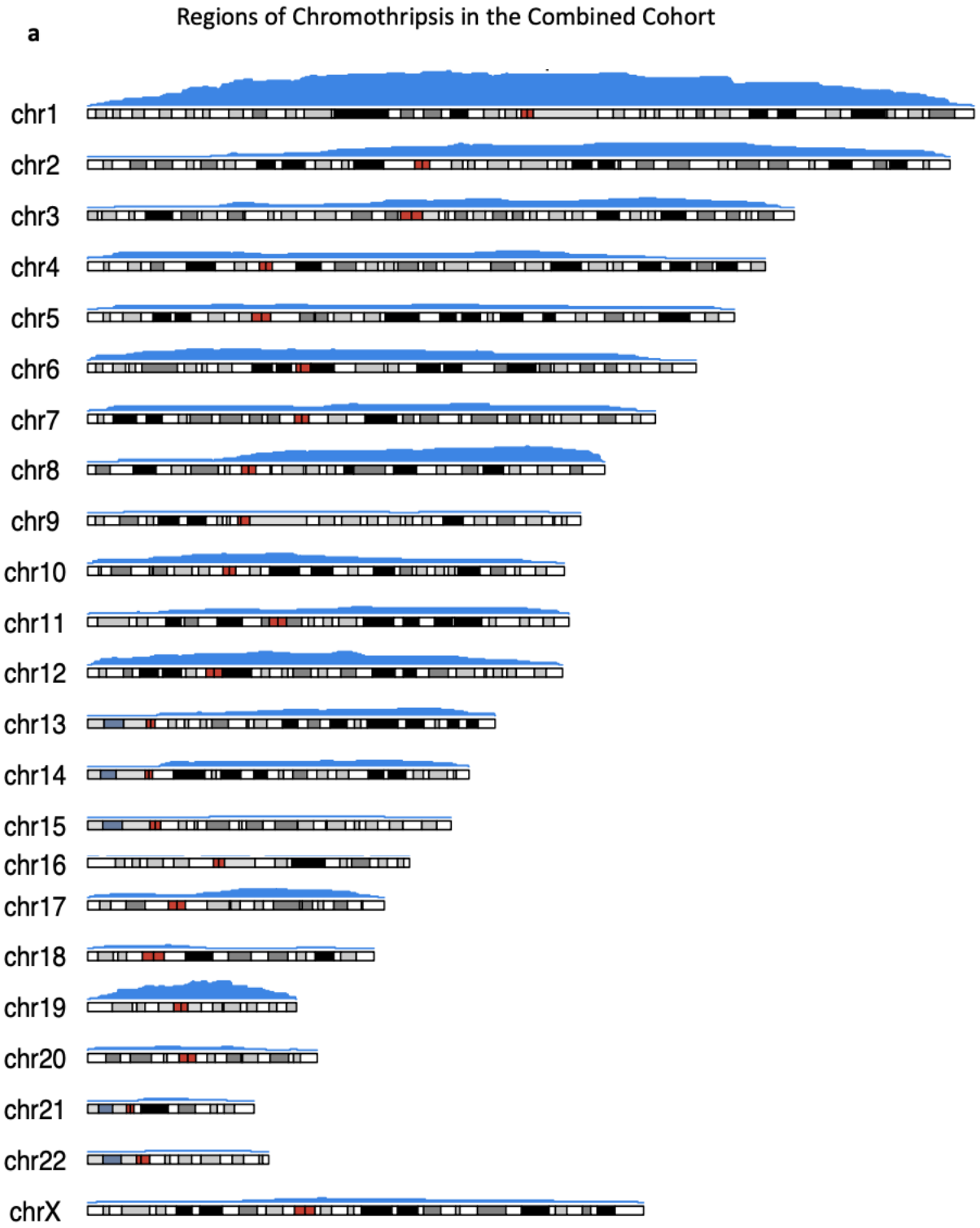


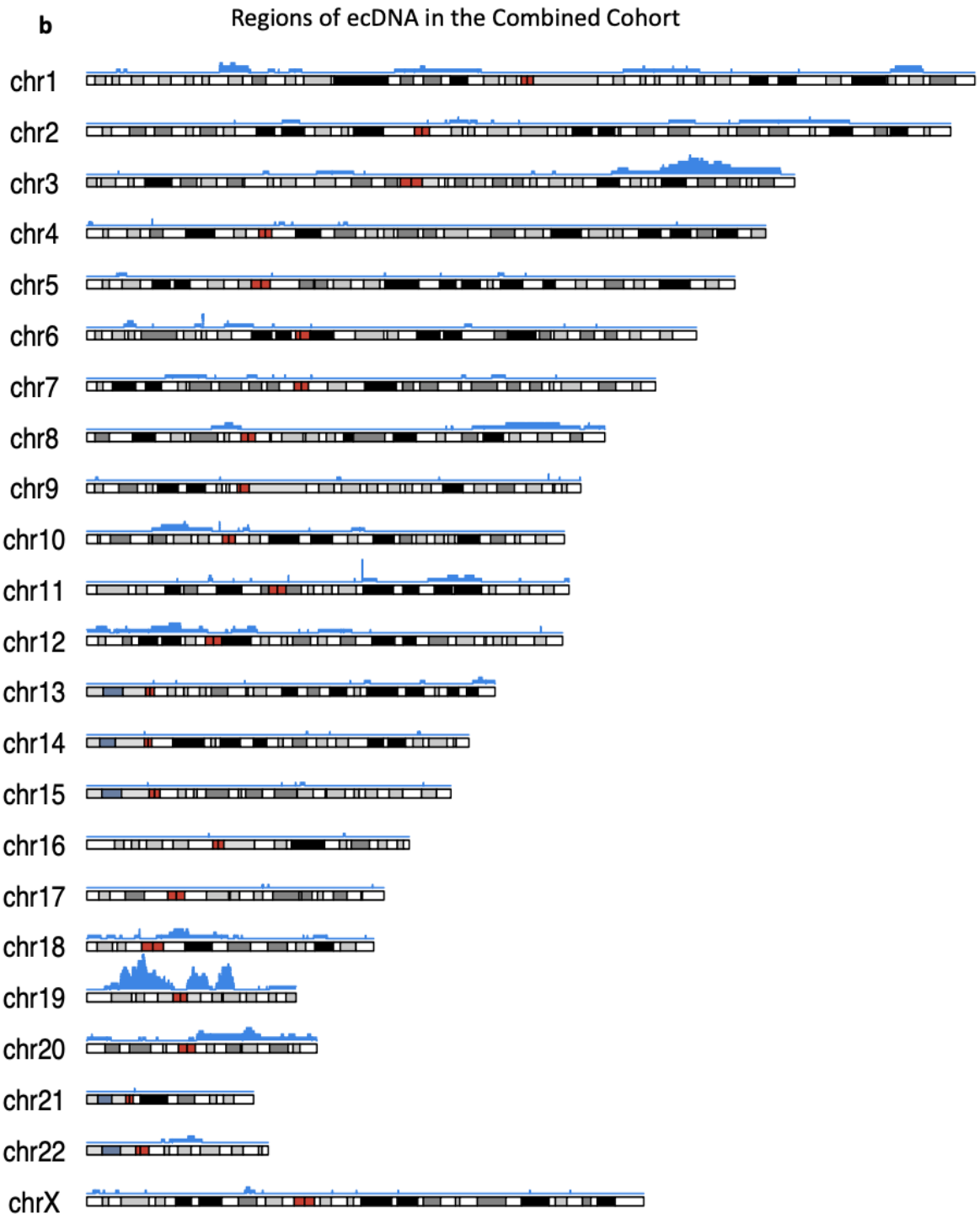
Figure 36 Correlated occurrence of SVs and CNVs within chromosomes

Log odds of SV and CNV occurrence per chromosome (as in Figure 34 and Figure 35) indicating enrichment (positive values) or depletion (negative).

Enrichment of Complex Structural Variants on Chromosomes

To investigate the distribution of cSV across chromosomes, the regions of each chromosome impacted by cSV across the combined cohort was plotted and shown in Figure 37. A build-up of chromothripsis on chromosome 1 and 19 (Figure 37a), additionally ecDNA (Figure 37b) and BFB (Figure 37c) show a build of cSV regions on chromosome 19. Notably, chromosome 16 did not display any regions of chromothripsis, as seen in Figure 37a. Pyrgo, rigma, tyfonas, and seismic amplification did not show a build-up of event on chromosome 19. However, it is challenging to determine from Figure 37 if the enrichment of cSV on chromosomes is significant.


a
Regions of Chromothripsis in the Combined Cohort
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX


b
Regions of ecDNA in the Combined Cohort
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX

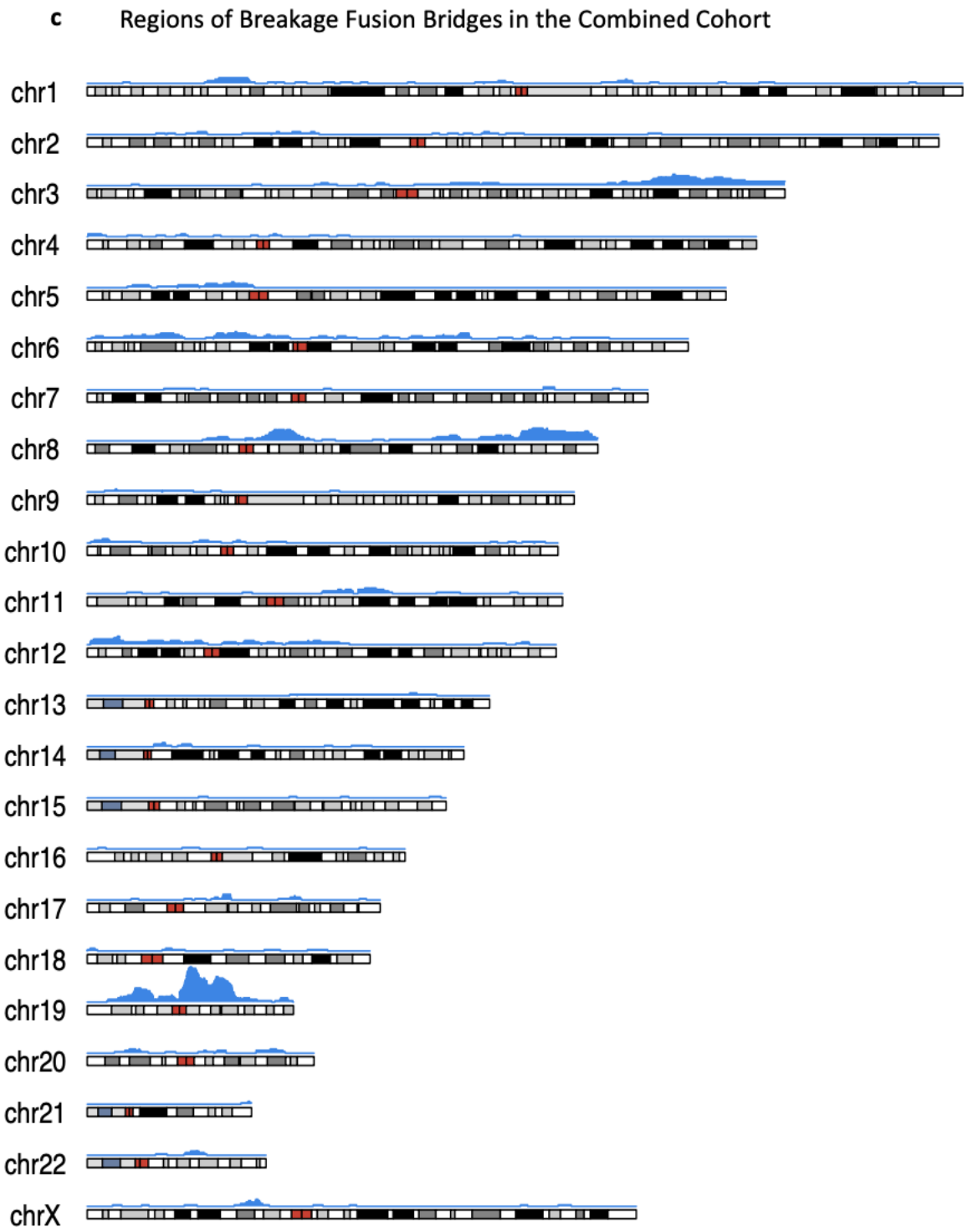


Figure 37 Evidence for cSV hotspots at multiple regions across the genome

For three complex structural variants (cSVs) types; chromothripsis, ecDNA and Breakage Fusion Bridge the region covered these cSV across each chromosome is shown in blue in **a**, **b** and **c** respectively. The height of peaks represent the number of samples where cSV regions overlap across the chromosome. The coverage is scaled so the most overlapped region fills the space between chromosomes.

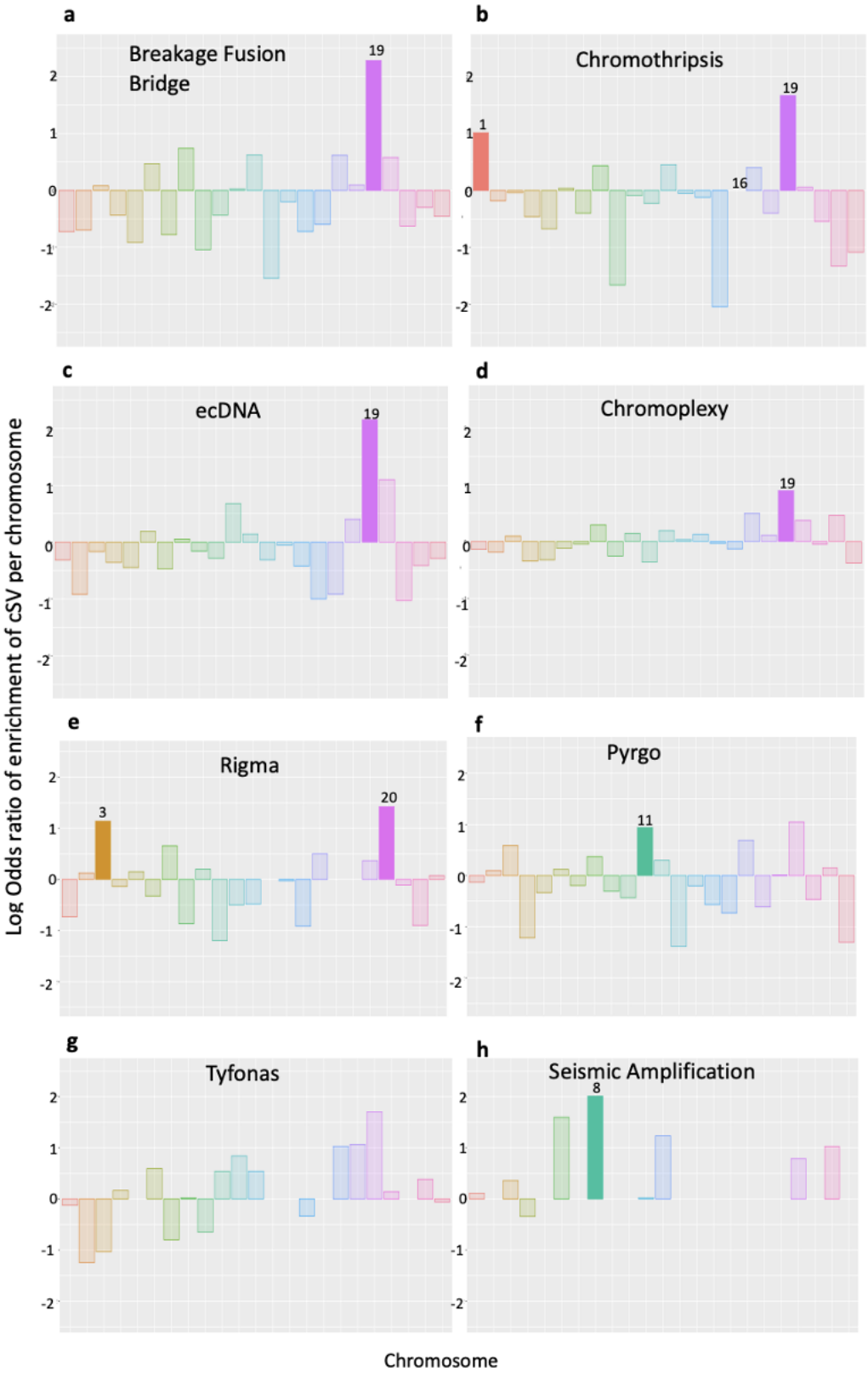


Figure 38 Not all complex structural variants are enriched on chromosome 19

Observed occurrence of cSVs across chromosomes relative to expectation, based on chromosome length, for (a) occurrence of breakage fusion bridge, (b) chromothripsis, (c),

ecDNA, (d), chromoplexI(e), ,rigma, (f), pyrgo, (g), tyfonas, (h), seismic amplification. Positive values (y-axis) of log odds ratios indicate relative enrichment and negative values indicate depletion. Significance was assessed using chi-squared tests ($p<0.05$ after Bonferroni correction) and is indicated by solid colours, translucent colours indicate a lack of significance.

To further investigate enrichment of cSV on chromosomes, the proportion of cSV on each chromosome was compared to the proportion of the genome that the chromosome comprises (Figure 38). Chromosome 19 shows significant enrichment for chromoplexy, chromothripsis, breakage fusion bridge (BFB) and ecDNA events. Their respective odd ratio shows their enrichment to be 2.5, 9.8, 5.3, 8.6 times more than would be expected based on chromosome length. However, not all types of cSV were enriched on chromosomes 19.

Pyrgo and rigma are most enriched on chromosome 20. Their respective odd ratio showing their enrichment to be 2.9 and 4.1 times what would be expected based on chromosome length shown in Figure 38. Additionally, pyrgo displayed enrichment on chromosomes 11 and 17, with an odd ratio of 2.6 and 2, respectively, suggesting that it was 2.6 and 2 times more likely to occur on those chromosomes than expected by chance.

Although tyfonas and seismic amplification do show enrichment, the low number of samples in which these events are identified limits the extent to which their enrichment is informative. Despite only being called in a few samples, the enrichment for seismic amplification on chromosome 8 did reach statistical significance.

Depletion of cSV on a chromosome can also be identified using the odds ratio. Although, depletion cannot be calculated on chromosomes where the cSV are never observed, it

can be estimated by comparing a chromosomes of a similar length with one cSV identified on it. For example in Figure 38, it is shown that chromosome 16 is never observed to have chromothripsis on it. However chromosome 15, which has a similar length to chromosome 16, has 0.1 times (or 10%) the amount of chromothripsis that would be expected by chance based on chromosome length. But, chromosome 15 does not reach statistical significance for depletion of chromothripsis.

In chapter 3 the enrichment of cSV in samples with WGD was discovered (Table 5). Samples with cSV impacting chromosome 19 were significantly enriched in samples with WGD and occurred more than twice as often as would be expected by chance (Table 9).

	Samples with WGD	Samples without WGD
Samples with cSV on chromosome 19	63	42
Samples without cSV on chromosome 19	79	140

Table 9 Whole genome doubling and complex structural variants impacting chromosome 19

The number of samples with and without whole genome duplications (WGD) and complex structural variants (cSV) impacting chromosome 19. A fisher exact test was used to test significance $p = 7.16x10^{-5}$, lower confidence interval 1.60, upper confidence interval 4.41 and odd ratio of 2.65.

Some sites within the genome are more susceptible to breakage than others, these fragile sights are not evenly distributed across the chromosome and could offer a potential explanation for the uneven distribution of cSVs (Kumar et al. 2019). By exploiting the HumCFS database of common fragile sites, the enrichment or depletion of fragile sites across chromosomes can be investigated (Kumar et al. 2019). Figure 39 a shows the enrichment and depletion of fragile sights across the chromosomes. The chromosome with the greatest significant enrichment of fragile sights is chromosome X followed by chromosome 19. It is tempting to speculate that chromosome 19 is enriched

for certain cSV types due to the enrichment of fragile sights however, chromosome X showed greater enrichment of fragile sites and did not show enrichment for any cSV type (Figure 38).

In Figure 39 b and Figure 39 c the enrichment of fragile sights is compared with the enrichment and depletion of CNVs and SVs. In both cases there was no trend between fragile sight and CNV or SV enrichment was observed except chromosome 19 which appears to be an outlier.

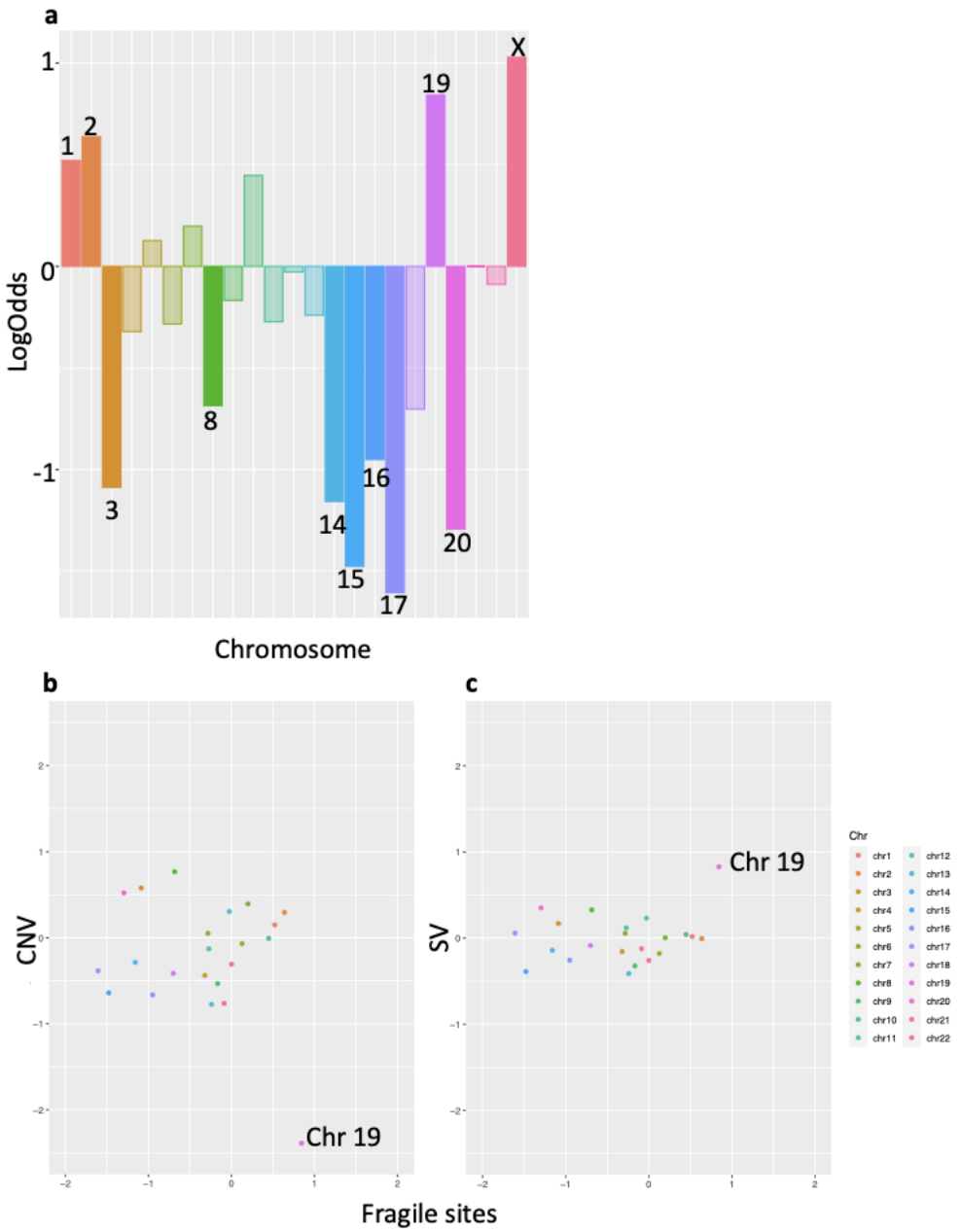


Figure 39 Enrichment of fragile sites across chromosomes

Observed occurrence for fragile sites from the curated database HUMcfs (Kumar et al. 2019) across chromosomes relative to expectation, based on chromosome length (a). The enrichment and depletion of chromosomes for fragile sites was tested using a Chi-squared test **a**. To more clearly show depletion, the log of the effect size (odds ratio) is used where positive values indicated enrichment and negative value indicate depletion. Significance is indicated by solid colours. The relationship between enrichment or depletion of fragile sights and enrichment or depletion of CNV **b.** The relationship between enrichment or depletion of fragile sights and enrichment or depletion of SV **c.**

Positive values indicate enrichment and negative values indicate depletion.

Figure 37 and Figure 38 have both shown an enrichment of chromothripsis on chromosome 19 and 1. It has been reported that larger chromosomes are more likely to be impacted by chromothripsis based on normalizelising single cell sequencing of multiple diploid cell lines and organoids including HeLa (Klaasen et al. 2022). However, it was not clear that this adequately accounted for the fact that the size of a chromothripsis region is proportionate to the size of the chromosome on which it occurs. Nevertheless, Figure 38 shows significant enrichment on chr 1 and 19, the largest chromosome and one of the smallest chromosomes respectively, which contradicts the overall correlation with chromosome size reported (Klaasen et al. 2022).

Closer examination of the known cSV events on chromosome 19 revealed a catastrophic landscape of combined intense SV and cSV activity (Figure 40) with focal enrichments of various SV classes and cSV events in the neighbourhood of the CCNE1 locus. It seems likely that the local increase in SV breakpoint densities seen at this locus is driven by increased chromothripsis, ecDNAs and BFBs (Figure 40). The gene, CCNE1 is a known oncogene in HGSOC, reported to be amplified (CN>=5) in~20% HGSOC cases and impact survival (Sapoznik et al. 2017). Importantly, the literature suggests that recurrent amplifications of CCNE1 drives tumourigenesis in HGSOC and this is also evident in the combined cohort (Figure 41). The amplification of CCNE1 (CN>=5) is found in 20.01% of samples and a fraction of these (43%) are associated with cSVs such as chromothripsis, ecDNAs and BFBs (Figure 41). It also appears that BFBs are associated with higher copy number amplifications of CCNE1 (Figure 41) which may be of clinical significance.

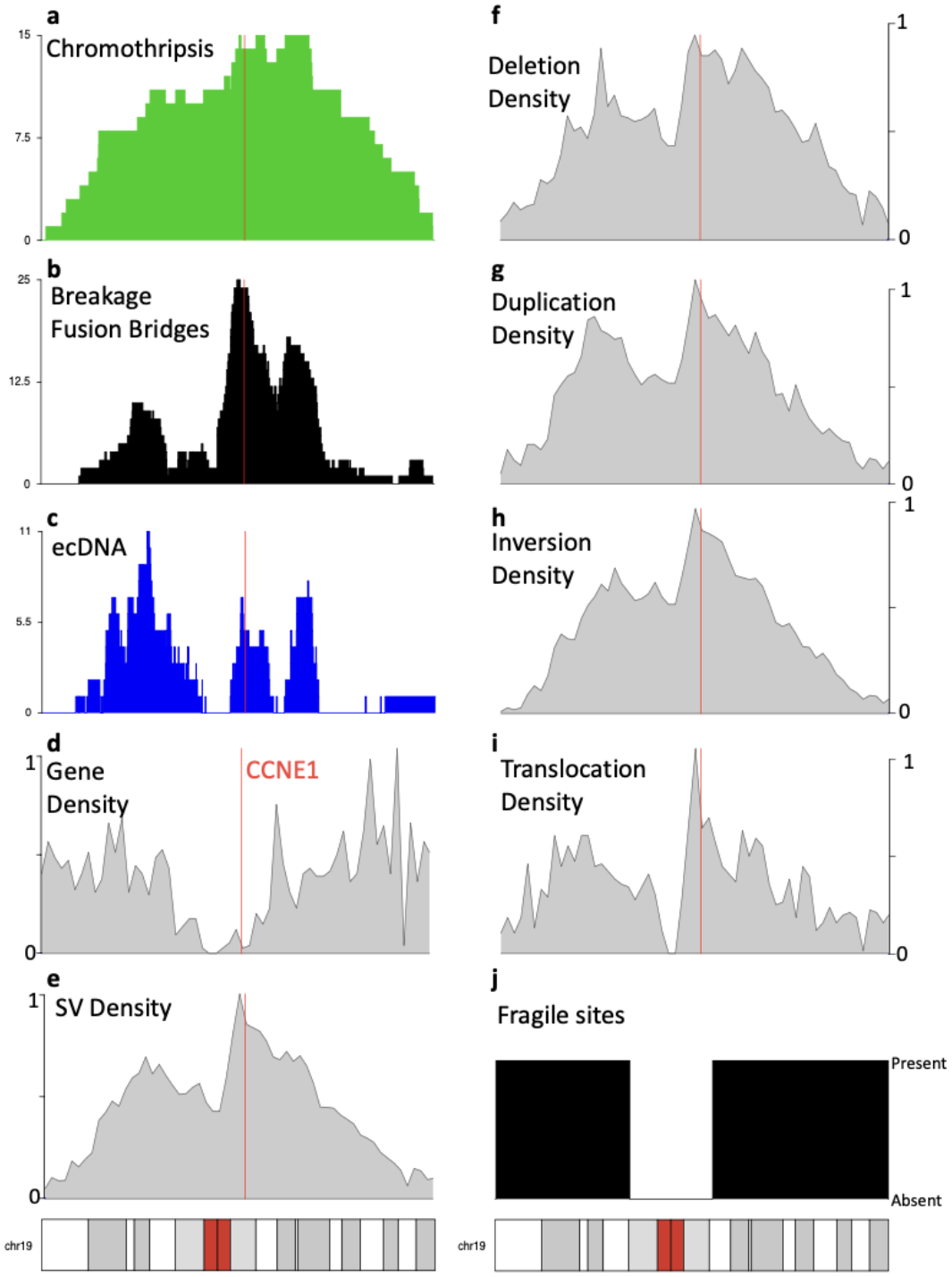


Figure 40 Alignment of complex structural variants and structural variants hotspots on chromosome 19

Hotspots of complex structural variants (cSV) and structural variants (SV) density across chromosome 19 with the red line indicating the location of known HGSOC oncogene CCNE1. Regions of chromosome 19 overlapped by chromothripsis are shown in green.

The number of samples in which chromothripsis overlaps a region, is shown by the height of the green peak in **a**. Regions of chromosome 19 overlapped by breakage fusion bridges are shown in black. The number of samples in which breakage fusion bridges overlaps a region, is shown by the height of the black peak in **b**. Regions of chromosome 19 overlapped by ecDNA are shown in blue. The number of samples in which ecDNA overlaps a region, is shown by the height of the blue peak in **c**. the density of genes across chromosome 19 is shown in **d.** The density of SV of all types across chromosome 19 is shown in **e**. The density of SV by each SV type is shown in **f**, **g**, **h** and **i** for deletions, duplications, inversions and translocations respectively where 1 represent maximum density on chromosome 19 and 0 represent the minimum density. The fragile sites located on chromosome 19 are shown in **j**. Density was calculated for each feature using non overlapping one megabase windows.

Figure 40 shows a peak of chromothripsis, BFB and ecDNA, total SV density as well as a peak of density for all SV types on or near the location of CCNE1- a gene widely reported to be amplified and important in HGSOC (Sapoznik et al. 2017; Petersen et al. 2020; Mei et al. 2023; Kroeger and Drapkin 2017; Raab et al. 2020). To investigate the link between copy number amplification of CCNE1 and cSV, the copy number of CCNE1 amplified by BFB was compared to simple amplification of CCNE1 in Figure 41 a. The amplification of CCNE1 by BFB was significantly greater than amplification of CCNE1 by other sources (Figure 41 a and c). Interestingly in the combined cohort ecDNA and chromothripsis only overlap amplified CCNE1 in the presences of BFB overlapping CCNE1 (Figure 41 b).

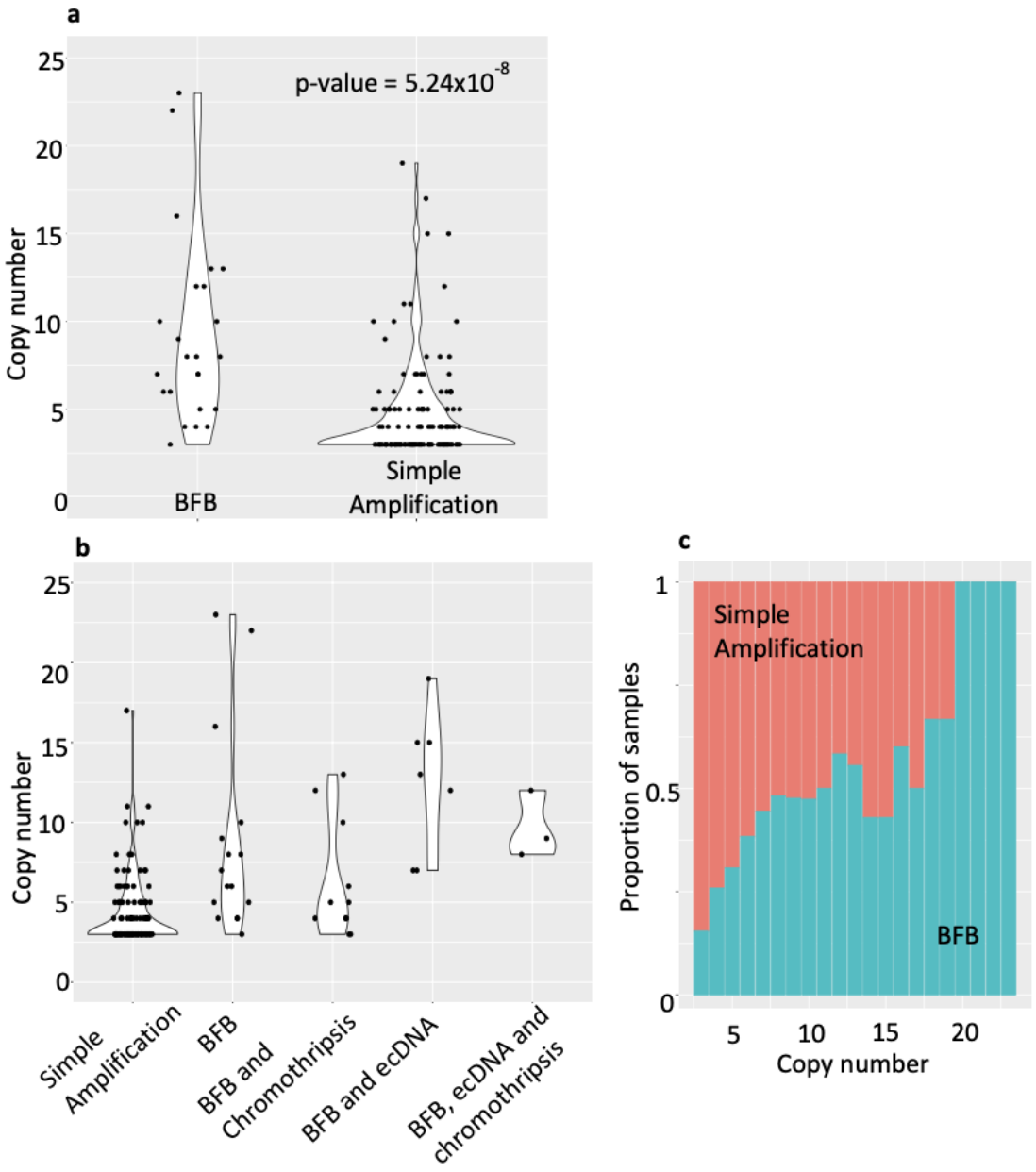


Figure 41 Amplification of CCNE1 by known cSV s

(2018) (a) The copy number of CCNE1 amplified by breakage fusion bridges (BFB; median CN=8, IQR=6.25) or simple (CN=3.5, IQR=2) amplification differs significantly (Wilcoxon test). (b) Copy number of CCNE1 amplified by any combination of cSV or simple CNV events. c) The proportion of samples possessing simple (red) or BFB (blue) mediated amplifications of CCNE1 is related to the resulting copy number of the amplification event.

When tested the overlap between BFB and CCNE1 occurred significantly (P = 0.0002) more than would be expected by chance (Figure 69 in Appendix).

Recurrent disruption of genes by cSV is common in HGSOC

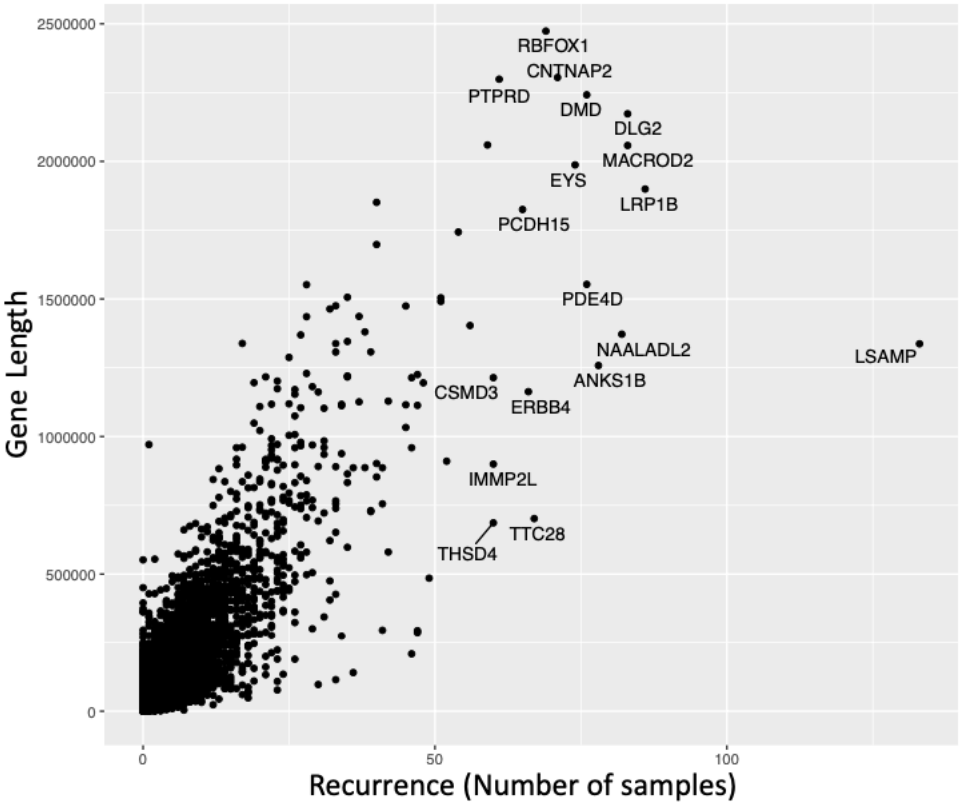
Before further investigating the impact of cSV on genes, the impact of simple SV on genes must be assessed. Across the combined cohort, how frequently a gene is disrupted was assessed by counting the number of samples where at least one breakpoint of an SV is located within the gene. This was done for all SV types combined and individually (Figure 42).

The most recurrently hit gene by all SV types (Figure 42 a) was LSAMP with 133 samples (39%) having at least one breakpoint within the gene. This was driven by deletion SVs where 113 samples had at least one breakpoint of a deletion within the gene (Figure 42 b). LSAMP has been proposed as a novel tumour suppressor in osteosarcomas and the re-expression LSAMP has been shown to inhibit tumour growth in ovarian cell lines (Barøy et al. 2014; Kresse et al. 2009).

A gene that is disrupted in fewer samples but is highly studied is ERBB4 (Figure 42 c) which was found to be disrupted in 66 (19%) samples and has previously been reported to have a breakpoint of unknown significance in the tumour of a patient with HGSOC (Nath et al. 2021). Additionally, the expression of ERBB4 has been reported to be greater in HGSOC tumours that are resistant to platinum treatment which will be further investigated in the next chapter (Saglam et al. 2017).

The recurrence of all SV types and deletions also highlights DMD gene which was impacted in 76 (22%) and 53 (15%) samples respectively. DMD has been reported to be a novel biomarker for ovarian cancer where it was suggested that DMD expression had a relevant role in the pathogenesis in HGSOC (Farinella et al. 2022).

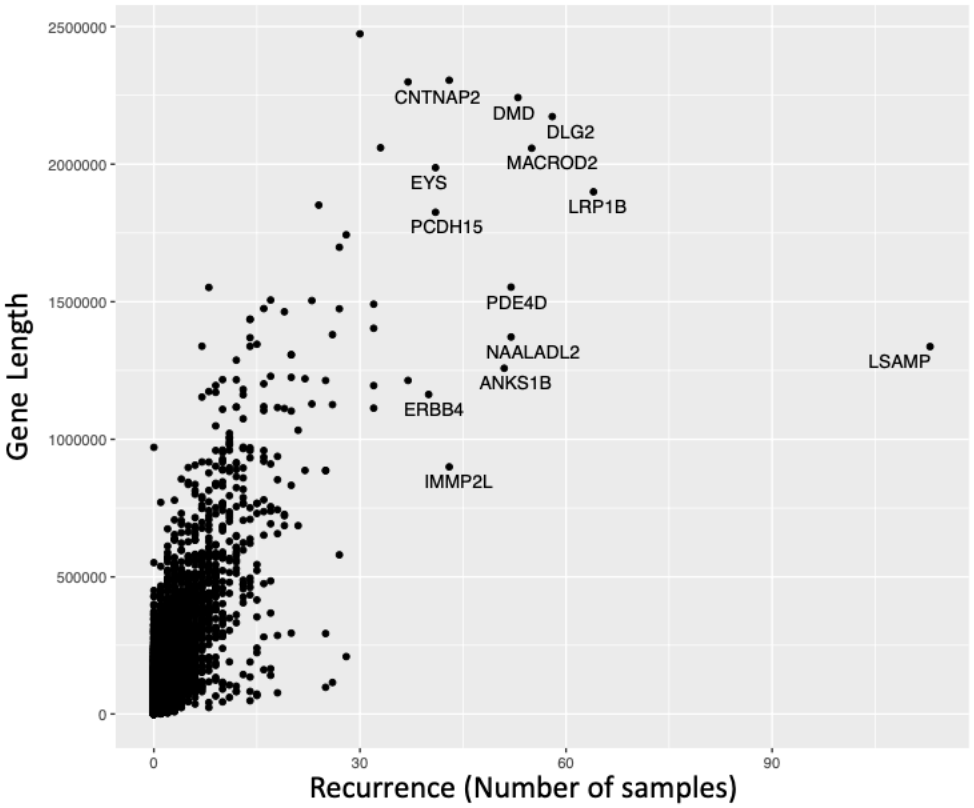
a i



a ii

Gene	Chr	Recurrence (All SV types)	Publications	Within Fragile sites
LSAMP	3	133	31	YES
LRP1B	2	86	225	YES
MACROD2	20	83	37	YES
DLG2	11	83	40	
NAALADL2	3	82	11	YES
ANKS1B	12	78	18	
PDE4D	5	76	99	YES
DMD	X	76	444	YES
EYS	6	74	126	
CNTNAP2	7	71	73	
RBFOX1	16	69	39	YES
TTC28	22	67	12	
ERBB4	2	66	1255	
PCDH15	10	65	15	
PTPRD	9	61	153	YES
THSD4	15	60	19	
CDMD3	8	60	61	
IMMP2L	7	60	14	YES

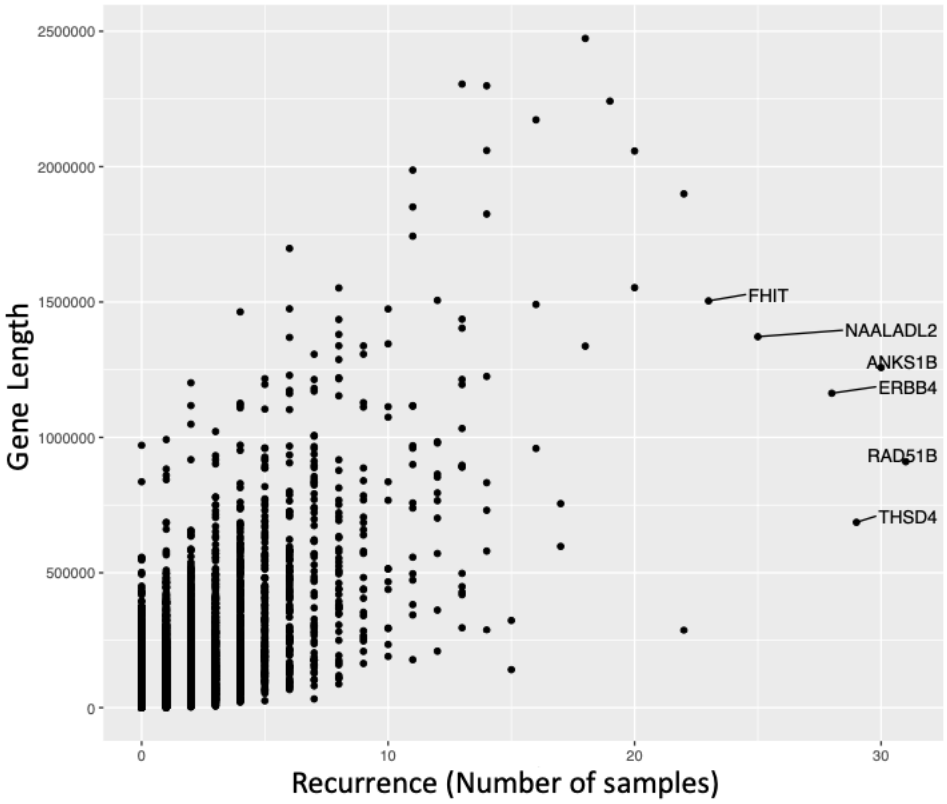
b i



b ii

Gene	Chr	Recurrence (Deletions)	Publications	Within Fragile sites
LSAMP	3	113	31	YES
LRP1B	2	64	225	YES
DLG2	11	58	40	
MACROD2	20	55	37	YES
DMD	X	53	444	YES
NAALADL2	3	52	11	YES
PDE4D	5	52	99	YES
ANKS1B	12	51	18	
CNTNAP2	7	43	73	
IMMP2L	7	43	14	YES
EYS	6	41	126	
PCDH15	10	41	15	
ERB4	2	40	125	

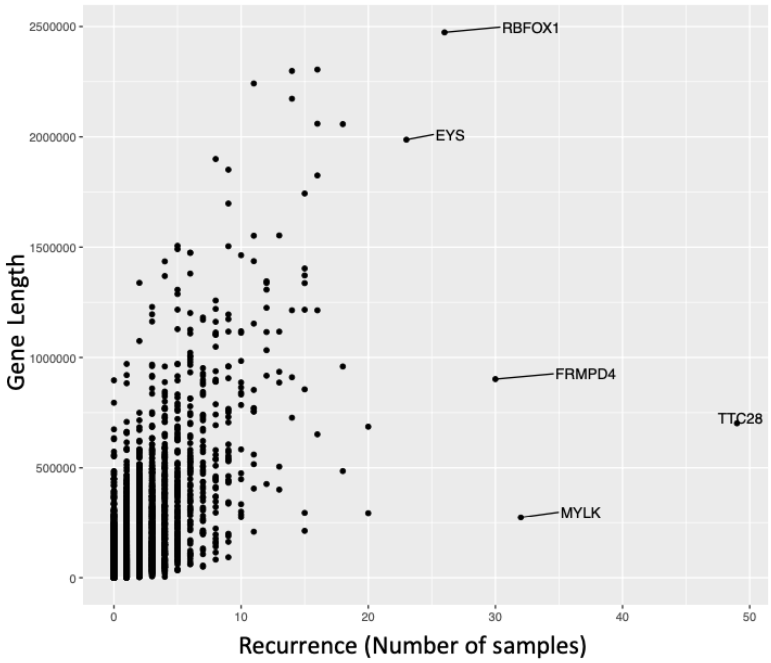
c i



c ii

Gene	Chr	Recurrence (Duplications)	Publications	Within Fragile sites
RAD51B	14	31	159	
ANKS1B	12	30	18	
THSD4	15	29	19	
ERBB4	2	28	1255	
NAALADL2	3	25	11	YES
FHIT	3	23	1107	YES
NF1	17	22	6190	
LRP1B	2	22	225	YES
MACROD2	20	20	37	YES
PDE4D	5	20	99	YES

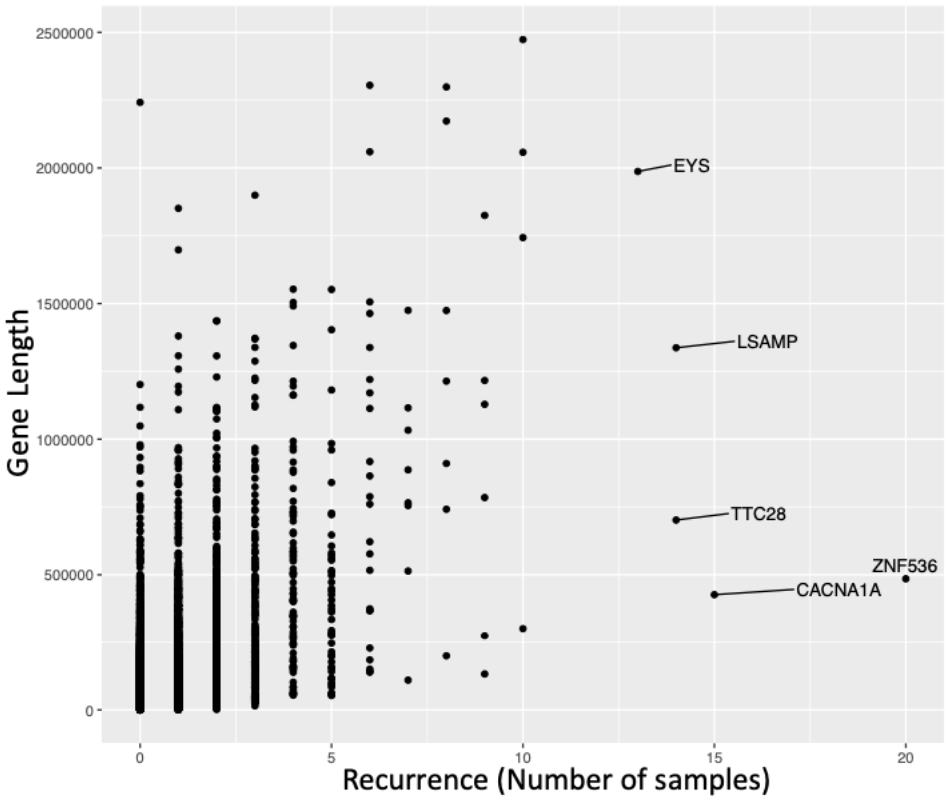
d i



d ii

Gene	Chr	Recurrence (Translocations)	Publications	Within Fragile sites
TTC28	22	49	12	
MYLK	3	32	90	
FRMPD4	X	30	2	
RBFOX1	16	26	39	YES
EYS	6	23	126	
EYA2	20	20	60	
THSD4	15	20	19	
DLGAP1	18	18	151	
ZNF536	19	18	1	
MACROD2	20	18	37	YES

e i



e ii

Gene	Chr	Recurrence (Inversions)	Publications	Within Fragile sites
ZNF536	19	20	1	
CACNA1A	19	15	40	

Figure 42 Genes recurrently disrupted by SVs

The number of samples in which at least one consensus structural variant (SV) breakpoint is within an exonic or intronic portion of a gene for each type of SV. A Spearman's correlations was used to test the correlation between recurrence and gene length for genes disrupted by all SV types Spearman's correlation = 0.74 p <2.20x10-16 (**a i**), genes disrupted by deletions Spearman's correlation = 0.62 p <2.20x10-16 (**b i**), genes disrupted by duplications (**c i**), genes disrupted by translocations Spearman's correlation = 0.57 p < 2.2e-16 (**d i**), genes disrupted by inversions Spearman's correlation =0.41 p < 2.2e-16 (**e i**). The most disrupted genes for each SV types are shown in the table below each graph (**a ii**, **b ii**, **c ii**, **d ii**, **e ii**). Additionally, the table shows the number of publications for results for the gene name and cancer found on Pubmed (September, 2022). Genes within fragile site is based on (Y. Li et al. 2020)

Recurrent duplications also impacted notable genes particularly NF1 which was disrupted in 22 (6%) samples. Interestedly, NF1 inactivation is thought to be beneficial to the tumour in HGSOC (Norris et al. 2018). Another gene impacted by duplications recurrently in 23 (7%) samples is FHIT. Similar, to NF1, FHIT disruption has been reported to be an important molecular event in HGSOC (Ozaki et al. 2001). This may suggest that disruption of a gene by a duplication breakpoint is as disruptive as by a deletion breakpoint.

However, it is important to note that not all genes that are recurrently disrupted will be tumour suppressor genes. Many will simply be genes which a tumour cell can tolerate being disrupted for examples LRP1B disrupted in 22 (6%) samples in the combined cohort has also been reported to be one of the most frequently mutated (by SNV) gene in ovarian cell line experiments (Yamulla et al. 2020). But, knocking out LRP1B did not increase transformation rates of the cell lines (Yamulla et al. 2020).

Understanding the consequence of a SV breakpoint being within the gene is an ongoing problem within the field (Weischenfeldt et al. 2013). A small disruption contained entirely within an intron may still yield functional protein yet a similar sized SV covering a exon may result in a non-functional protein. To investigate if recurrence was driven by SVs which were unlikely to impact the protein, the SVs impacting LSAMP were assessed Figure 43.

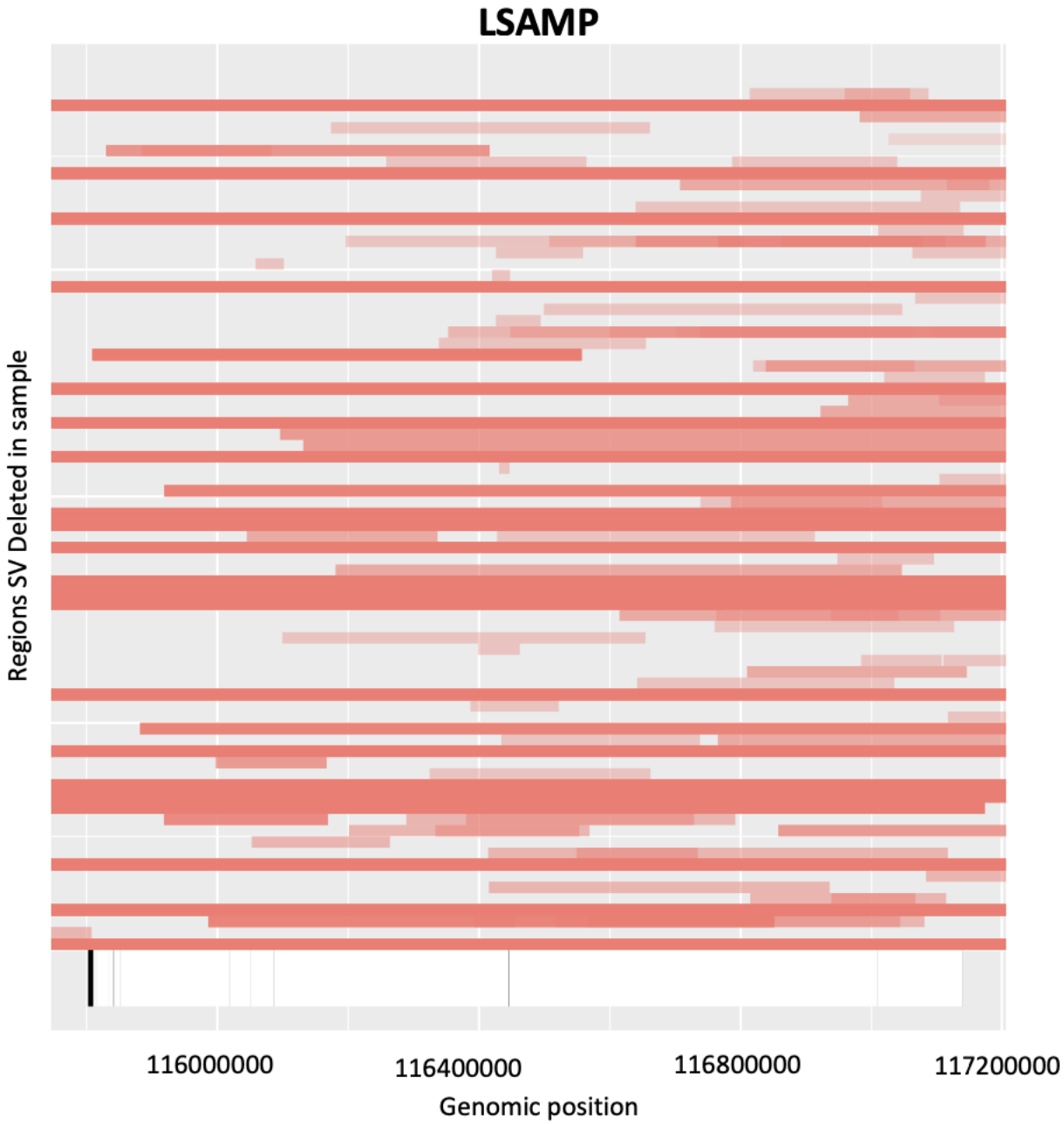


Figure 43 Many SV deletions impact LSAMP exons

There are 634 breakpoints within the LSAMP gene from 133 samples and only 2 samples had a breakpoint within the exon of the LSAMP gene. 76 samples had a consensus SV deletions spanning an exon of LSAMP. The gene shown at the bottom with exons are shown in black in the gene with introns in white. SV deletions are shown in red one sample per row.

Only 2 samples had a breakpoint within in LSAMP exons. Out of all samples, only 133 samples had at least one breakpoint within the introns of LSAMP which makes up ~ 99 % of the genes total length. In 76 samples deletions span at least one exon making it highly likely to be disruptive.

The vast majority of Genes are overlapped by Complex Structural Variants

As shown in Figure 44, the majority (84%) of protein coding genes were covered at least once by a cSV in the combined cohort. This means that less than 4000 genes were never covered by a cSV. Chromothripsis covered the most genes of any cSV type, covering 76% of protein coding genes.

As cSV regions can be megabases in length, some genes may be completely covered by a cSV without being directly impacted by a breakpoint. To identify the gene most likely to be disrupted by cSV, only genes coved by a cSV at least once and containing a breakpoint within the gene were decribed as disrupted Figure 44c. The number of genes disrupted by cSV is much smaller than the number of genes overlapped by cSvs.

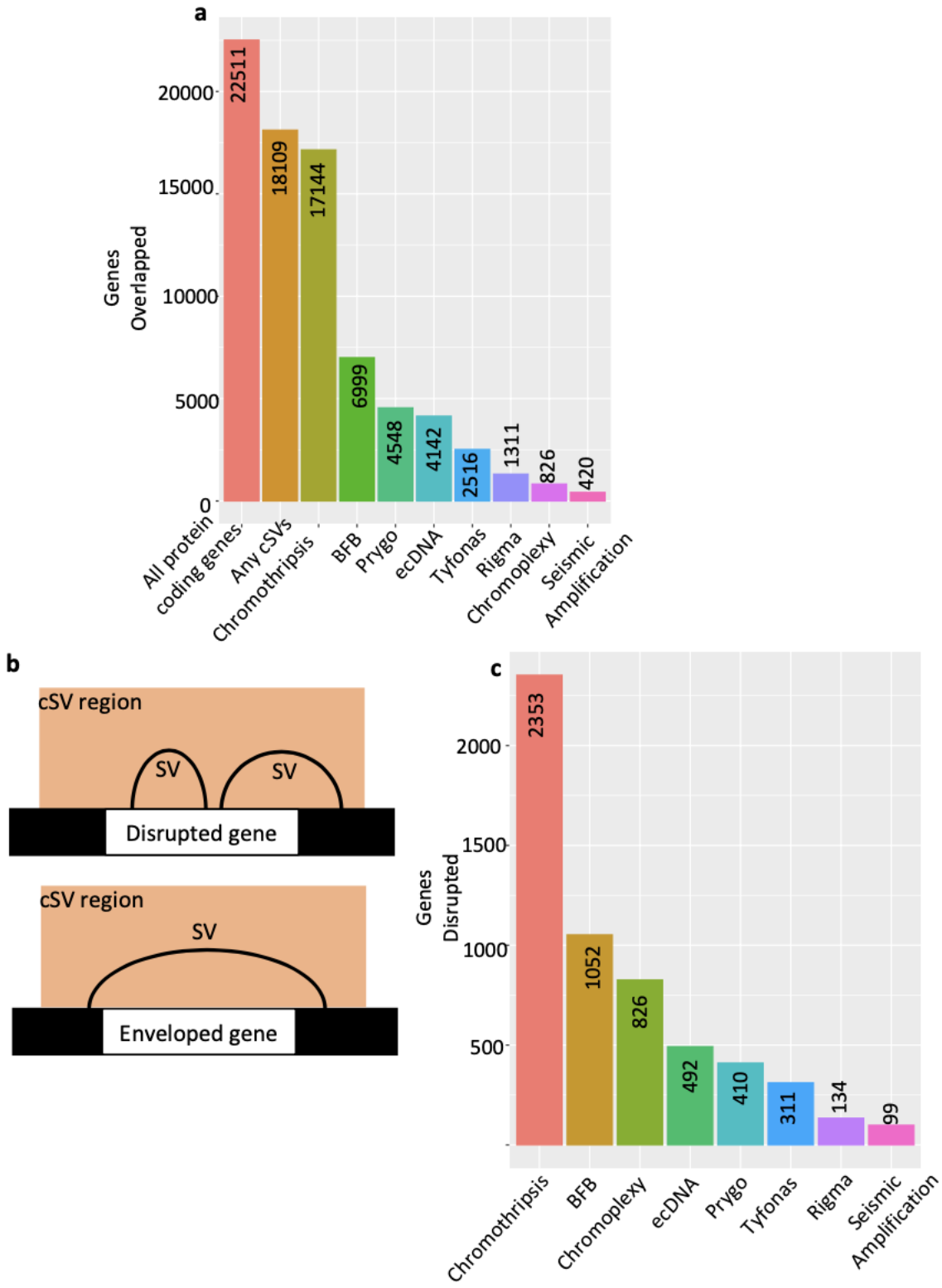


Figure 44 Genes affected by complex structural variant events

(a) The number of protein coding genes encompassed within a cSV in at least one sample is shown for all cSV types relative to all protein coding genes overlapped by any cSV at

least once. (b) Schematic for affected genes, including ‘disrupted’ genes which are interrupted by at least one SV breakpoint within the span of the gene, and ‘enveloped’ genes which are completely encompassed by a cSV but do not contain a SV breakpoint. (c) The number of protein coding genes directly disrupted by cSVs in at least one sample is shown for all cSV types.

The Impact of cSVs on Clinically Relevant Genes

From the curated list of genes thought to be of clinical relevance in HGSOC, the number of samples in which each gene is overlapped by a cSV to any extent was identified (Figure 45) (Hollis et al. 2022). The majority of cSV types do not overlap clinically relevant genes. However, certain genes were notable: BFBs commonly overlapped CCNE1 (24 samples) and chromothripsis often overlapped RAD54L (20 samples). Beyond this, it is notable that most genes in this list were overlapped by chromothripsis to some degree across the cohort.

The cSV gene pair that overlapped the most was BFB overlapping CCNE1 in 24 samples. Additionally, CCNE1 was overlapped by ecDNA in 7 samples and amplification of CCNE1 by ecDNA has been reported to be important in tumour development (X.-K. Zhao et al. 2021a).

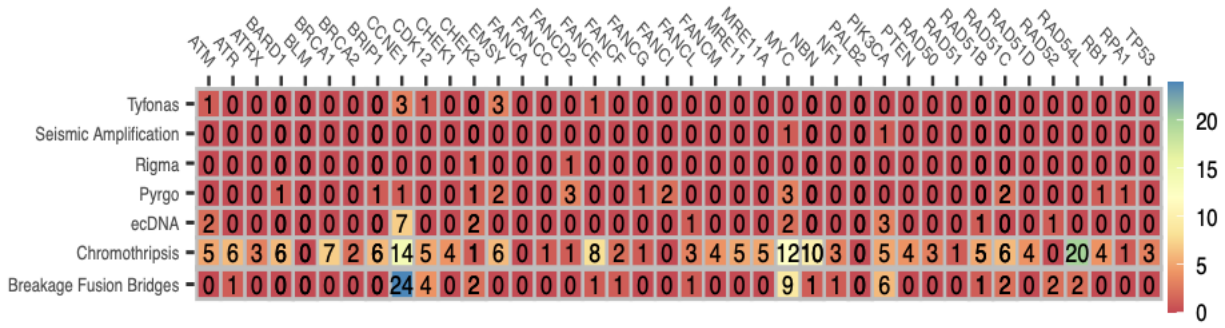
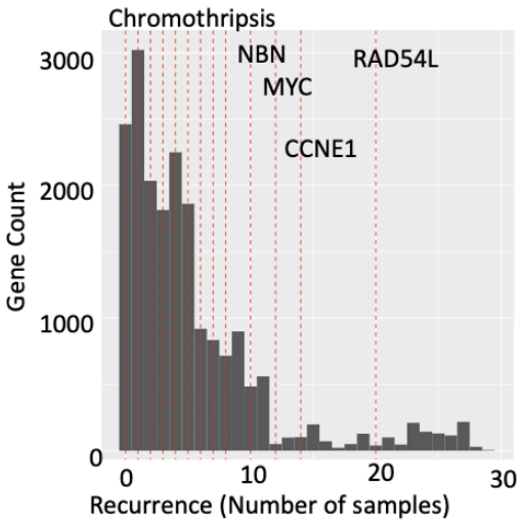


Figure 45 Clinically relevant genes and complex structural variant events.
The number of samples where each cSV type overlaps clinically relevant genes in the combined cohort. The clinically relevant genes where selected based on their reported importance in HGSOC (Hollis et al. 2022). The cells are colored from red for low overlap

to blue for high overlap between cSV and Genes.

To compare the overlap of cSV with clinically relevant genes and selected genes for further analysis by circular permutations the recurrence of overlap between cSV and all protein coding genes was assessed Figure 46. CCEN1 is one of the most recurrent genes impacted by BFB being overlapped in 24 (7%) samples Figure 46 c and d. The DNA repair gene, RAD54L, was recurrently overlapped by chromothripsis in 20 (6%) samples. MYC was overlapped by BFB in 9 (2%) samples the significance of this is hard to asses so circular permutation, which allows for random sampling of regions across a chromosome, will be used for further investigation

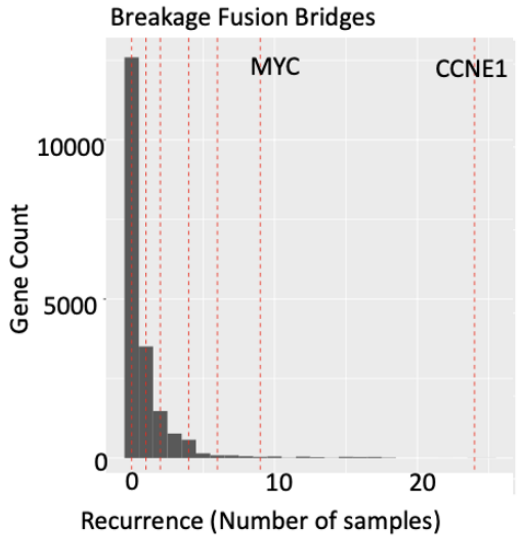
a



b

Gene Name	Chr	Recurrence	Publications
S1PR1	1	29	239
DPH5	1	29	3
SLC30A7	1	29	9
EXTL2	1	29	15
OLFM3	1	29	3
SLC44A3	1	28	4
ABCD3	1	28	28
ALG14	1	28	3
CNN3	1	28	25
DPYD	1	28	492
TLCD4	1	28	0

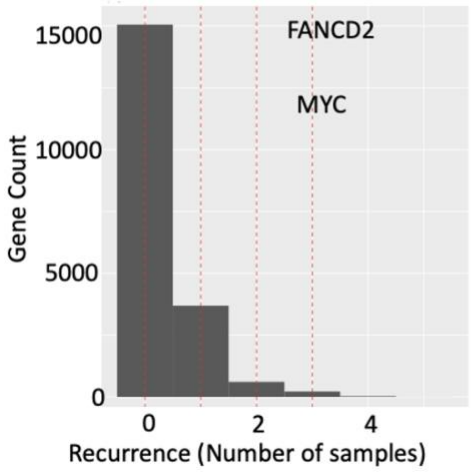
c



d

Gene Name	Chr	Recurrence	Publications
UQCRFS1	19	25	20
POP4	19	25	9
VSTM2B	19	25	2
PLEKHF1	19	25	9
ZNF536	19	24	1
URI1	19	24	43
CCNE1	19	24	814
C19orf12	19	24	3
TSHZ3	19	22	13
ZNNF790	19	18	0
ZNF345	19	18	0
ZNF345	19	18	2

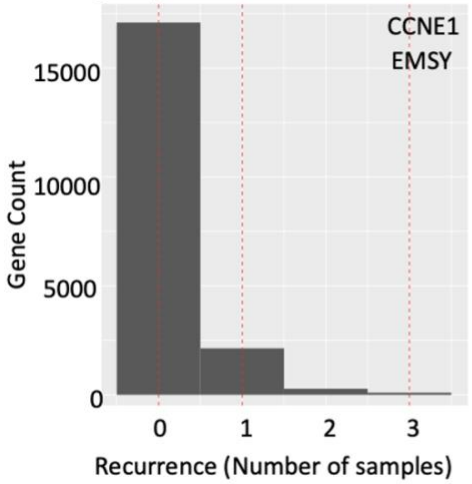
e Pyrgo



f

Gene Name	Chr	Recurrence	Publications
MECOM	3	5	459
GOLIM4	3	5	5
SERPINI1	3	5	21
PRRG4	11	4	11
EIF3M	11	4	13
WT1	11	4	3866
DEPDC7	11	4	3
TCP11L1	11	4	0
CSTF3	11	4	5
MYL4	17	4	5
EFCAB13	17	4	2

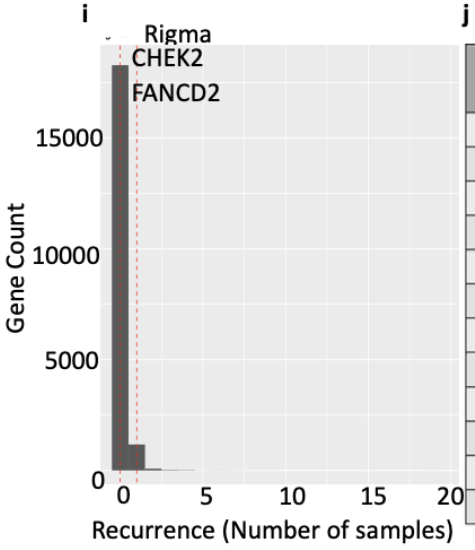
g Tyfonas



h

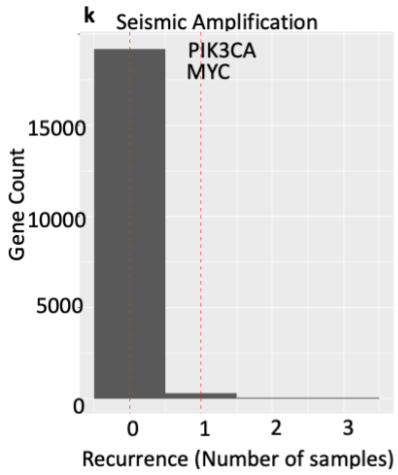
Gene Name	Chr	Recurrence	Publications
CCNE1	19	3	814
EMSY	11	3	74
NRP1	10	3	769
KIF5B	10	3	223
USP35	11	3	12
NARS2	11	3	3
CEBPG	19	3	19
FAAP24	19	3	24
TSHZ3	19	3	13
CHST8	19	3	2
CEBPA	19	3	1070
GAB2	19	3	231

i



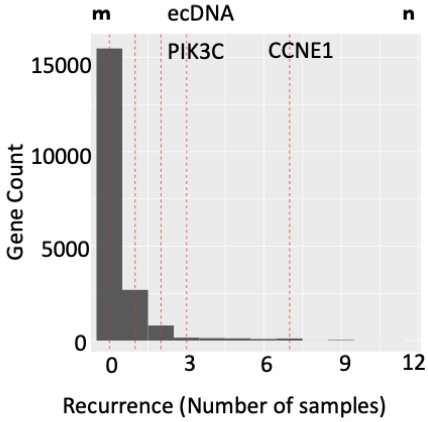
j

Gene Name	Chr	Recurrence	Publications
LSAMP	3	19	32
GAP43	3	15	209
MACROD2	20	9	37
FLRT3	20	8	13
SEL1L2	20	7	0
IMMP2L	7	7	14
LRP1B	2	7	227
ZBTB20	3	7	51
LINGO2	9	6	6
VAT1L	16	6	1
NDUFAF5	20	6	5
ESF1	20	6	2



l

Gene Name	Chr	Recurrence	Publications
CYP7B1	8	3	43
CA8	8	3	30
CLVS1	8	3	1
ARMC1	8	3	2
XKR9	8	3	2
ASPH	8	3	87
SULF1	8	3	141
C8orf34	8	3	5
STAU2	8	3	12
PRDM14	8	3	90
LY96	8	3	64
TRPA1	8	3	218



n

Gene Name	Chr	Recurrence	Publications
CYP4F2	19	11	60
CYP4F12	19	11	14
OR10H4	19	11	0
OR10H5	19	11	1
OR10H1	19	11	2
CYP4F11	19	11	12
RAB8A	19	10	54
TPM4	19	10	75
HSH2D	19	10	6
OR10H2	19	10	0
OR10H3	19	10	1
OR7C1	19	9	1

Figure 46 Clinically relevant genes show unusual patterns of cSV recurrence

Recurrence of cSV events in clinically relevant HGSOC genes of interest (red lines) relative to the recurrence seen in all protein-coding genes. (a,b) The recurrence of chromothripsis at all protein coding genes with outlier HGSOC genes of interest indicated (c,d) The recurrence of BFBs at all protein coding genes with outlier HGSOC genes of interest indicated. (e,f) The recurrence of pyrgo at all protein coding genes with outlier HGSOC genes of interest indicated. (g,h) The recurrence of typhonas at all protein coding genes with outlier HGSOC genes of interest indicated. (i,j) The recurrence of rigma at all protein coding genes with outlier HGSOC genes of interest indicated. (k,l) The recurrence of seismic amplification at all protein coding genes with outlier HGSOC genes of interest indicated. (m,n) The recurrence of ecDNA at all protein coding genes with outlier HGSOC genes of interest indicated.

Enrichment of Pan-cancer Gene Lists in cSV

To identify the genes that are most likely to be affected by cSVs, I narrowed down the disrupted gene list by only including genes that have some overlap with a cSV and at least one breakpoint within the gene region. As illustrated in Figure 46 c, this resulted in a reduction of the number of genes present in the disrupted gene list.

Due to the size of some cSV it is possible for a gene to be completely overlapped by a cSV and yet contain no breakpoints within the genes. The genes that had any overlap with a cSV and no breakpoints within the genes were collated into the enveloped gene list. As can be seen from Figure 44 fewer than 4000 protein coding genes were never hit by any cSV in any sample in the combined cohort. These genes were collated into the elusive gene list.

Beyond known HGSOC genes, there are a variety of other curated pan-cancer gene lists of interest. I examined the COSMIC curated cancer gene census (CGC) list and a dataset of genes found to be essential in cell lines for enrichment of cSV events (Sondka et al. 2018; Vinceti et al. 2021). A number (N=578) of oncogenes, cancer fusion genes and, tumour suppressor genes were defined by CGC annotation, while (N=542) essential genes were determined by the results of CRISPR-Cas9 screens from the DepMap database (Vinceti et al. 2021; Sondka et al. 2018). The genes examined were subdivided into disrupted genes (containing a SV breakpoint), enveloped genes (encompassed by cSV but not containing a SV breakpoint) and elusive genes (not impacted by cSV in any sample) (Figure 47).

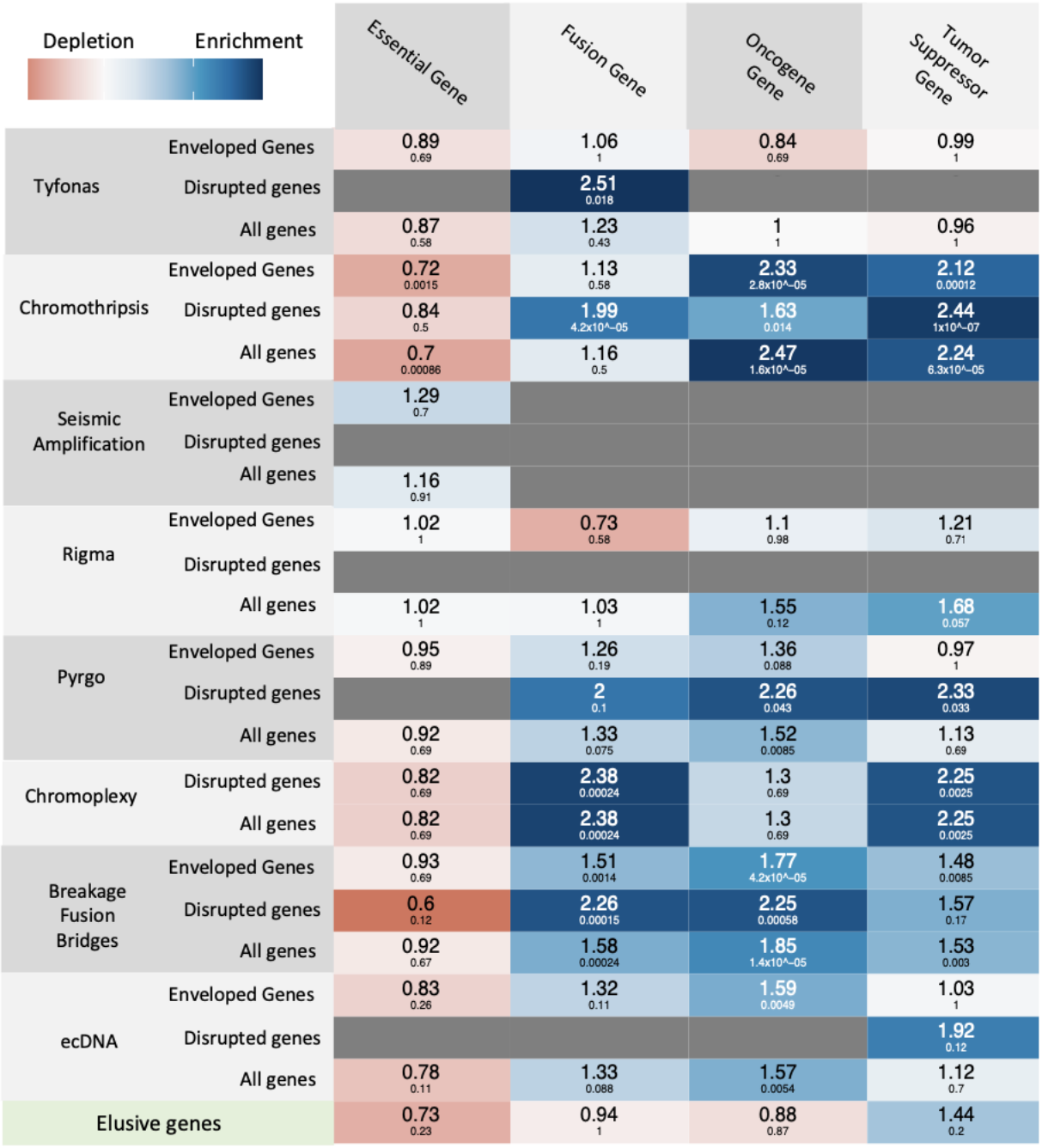


Figure 47 Pan-cancer gene lists and complex structural variant types

Heatmap of the enrichment or depletion of disrupted/enveloped pan-cancer CGC genes and DepMap essential genes with enrichment/depletion indicated by cell colour and the large number. Within each cell, the odds ratio is indicated with the (chi-squared) significance of enrichment/depletion for genes from that list in the smaller number. Only combinations with at least ten genes shared between the cSV and gene list were tested.

For the essential genes the only significant result is that regions of chromothripsis show modest but significant depletion for essential genes. This implies a weak bias against the disruption of essential genes by chromothripsis.

The only significant result for ecDNA was for enveloped oncogenes. Meaning these genes are highly likely to be functional and amplified by the ecDNA. No pattern was found between the elusive genes and pan-cancer gene lists, including the essential genes. This is consistent with the result for chromothripsis, in that any selection to preserve essential genes appears to be weak. Further analysis of the impact of these cSV on genes and subsequent impact on survival will be done in the following chapter.

Discussion

In this chapter, my objective was to explore the patterns and distribution of cSVs throughout the genome. To achieve this goal, I analysed 324 samples of HGSOC. Through this analysis, I was able to study the distribution of multiple types of cSVs across the genome and identify trends that might have been overlooked in prior studies due to smaller sample sizes and fewer identified cSVs.

This chapter has shown that cSVs are not uniformly distributed across chromosomes, with some chromosomes displaying higher levels of cSV enrichment than others. The breadth of cSV assessed allowed chromosome 19 to emerge as a hotspot of multiple cSVs. PCAWG reported enrichment of chromothripsis on chromosomes 3 and 5 in kidney renal cell carcinomas and chromosome 12 in liposarcomas, which had been previously reported (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020; Garsed et al. 2014; Mitchell et al. 2018). However, it is unclear how enrichment was calculated and if chromosome length was taken into account by PCAWG (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020). PCAWG did not report any enrichment of chromothripsis on chromosome 19. The investigations into ecDNA in the PCAWG dataset focused on enrichment of oncogenes but did not report enrichment at a chromosomal level (Kim et al. 2020). Because, the combined cohort contained nearly three times as many samples of HGSOC (n=324) than PCAWG (n=109) and because a wider range of cSV have been investigated, the pattern of a hotspot of cSV on chromosome 19 could be identified in this work.

The presence of fragile sites on chromosome 19 does not fully explain the result as although chromosome 19 was one of the few chromosome enriched for fragile sites, the chromosomes with the greatest enrichment for fragile sites was chromosome X which did not show the same enrichment for multiple cSV types. A study in 2015 revealed that fragile sites in prostate cancer were not associated with an increased incidence of chromothripsis (Kovtun et al. 2015).

Chromothripsis , BFB and ecDNA all had peaks of recurrence on or close to the known oncogene CCNE1 on chromosome 19. It was demonstrated that SVs and CNVs were unevenly distributed across chromosomes, with chromosome 19 being an outlier. Specifically, chromosome 19 was found to be enriched for SVs but depleted for CNVs.

In addition to this, enrichment of multiple cSV types on chromosome 19 where CCNE1 is present and overlapping the oncogene loci were observed. Amplification of CCNE1 has been proposed to be a key driver gene in HGSOC tumour development (Raab et al. 2020; Petersen et al. 2020; Gorski, Ueland, and Kolesar 2020). For the first time BFB has been shown to amplify the copy number of CCNE1 more than simple amplification. Moreover, while other types of cSVs types such as, ecDNA have previously been linked to the amplification of CCNE1, it was only observed that these cSVs overlapped with CCNE1 only when BFB events also overlapped with CCNE1 in the combined cohort. (X.-K. Zhao et al. 2021b; Turner et al. 2017; Nikolaev et al. 2014).

The impact on survival of the amplification of CCNE1 will be investigated in the next chapter. Additionally, the survival and gene expression effects of all the cSV identified later in this work.

Chapter 5: Linking Complex Structural Variation to Gene Expression and Survival

Introduction

In the previous chapters, complex structural variants (cSVs) were shown to have a non-random distribution across sub-cohorts, chromosomes, and to occur at high frequencies at the locus of the known oncogene CCNE1. In this chapter, I will compare samples with and without cSVs to assess the effects of cSV types on gene expression and patient survival time after diagnosis.

The impact of cSVs on gene expression has been inconsistently studied across cSV types, since few studies have examined more than a single type. Many cSV types, such as ecDNA, chromothripsis, BFB, and seismic amplification, have been reported to increase gene expression of individual genes in some samples. (Z. Li et al. 2022; Shoshani et al. 2020; Bianchi et al. 2019; Rosswog et al. 2021). In this chapter, I will investigate the broad effect of eight cSV types on the expression of all protein coding genes using consistent, statistically rigorous methods.

The reported impact of cSV on survival varies by cSV type and cancer type, and as with studies of gene expression, studies of survival have adopted different methods in varying collections of tumour types. Studies of ecDNA encompassing many tumour types have reported worse survival for patients with tumours containing ecDNA (Adelman and Martin 2021; Kim et al. 2020; Robert and Crasta 2022; Shah et al. 2021; Bailey et al. 2020), but the survival impacts of most other cSV types are less well studied. Studies of chromothripsis have been contradictory, with reports of associations of better and worse survival (Fontana et al. 2018; Forment, Kaidi, and Jackson 2012; Kloosterman, Koster, and Molenaar 2014; Magrangeas et al. 2011; Molenaar et al. 2012; Skuja et al. 2017). Furthermore, it should be noted that cSV of the same type are usually far from identical and exist on a severity spectrum, related to the size of the events and the number of SVs involved. The survival impacts of cSV across a range of severity is understudied, but will be explored in this chapter.

The key questions to be addressed in this chapter are:

i. Survival: Does the presence of different complex structural variant types found in the combined cohort have a consistent impact on survival, or are some types associated with better or worse outcomes?

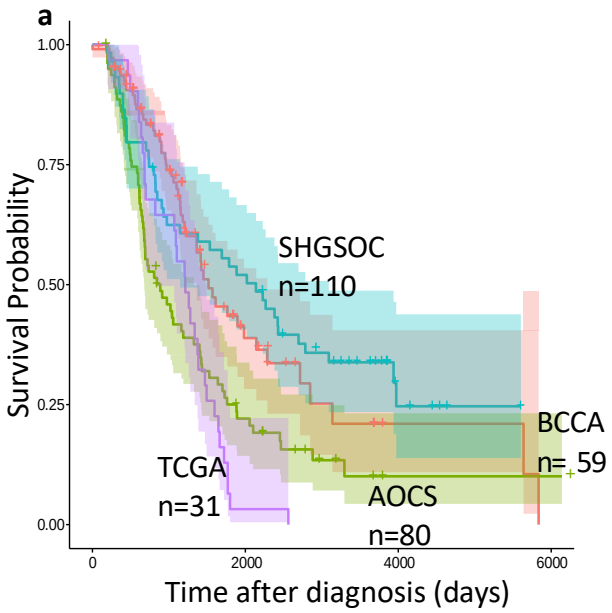
ii. Gene Expression: Do complex structural variant types consistently alter gene expression, and are there differences in these alterations depending on the type of the complex structural variant?

iii. Variation of Severity: How does the severity of complex structural variant events impact survival?

Differences in Survival between the Four Sub-cohorts

Before investigating the effect of cSV on survival and gene expression the differences on survival in the four sub-cohorts was assessed. To investigate if there are any differences in survival between these sub-cohorts, the survival time after diagnosis was viualised using a Kaplan-Meier survival curve (Figure 48a) and tested using a Cox proportional hazard model adjusting for age at diagnosis, stage at diagnosis, homologous recombination deficiency (HRD) status, and whole genome duplication (WGD) (Figure 48 b). The survival curves show that the TCGA and AOCS sub-cohorts have worse survival than the BCCA and SHGSOC sub-cohorts (Figure 48a). The AOCS and TCGA sub-cohorts have a significantly increased risk of death (P= 0.003, HR= 1.79 and P= 0.027, HR= 1.67 respectively) compared to the SHGSOC (Figure 48 b). This shows that there is a significant effect of sub-cohort membership on survival in the combined cohort and so this will be adjusted for in subsequent analyses.

a



b

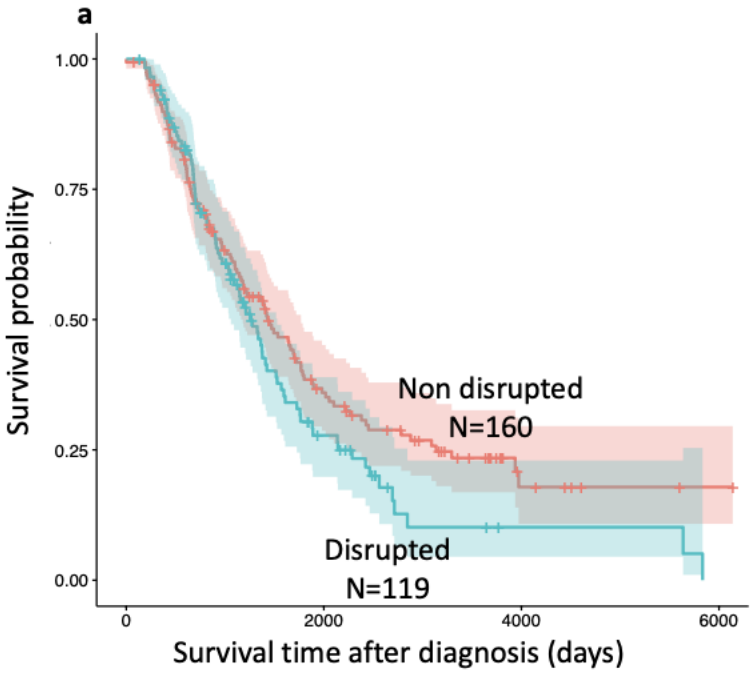
Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-103)	0.07	
Stage	N=280	1.54 (1.17-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-1.63)	<0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.93 (0.69-1.25)	0.627	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.79 (1.22-2.62)	0.003	
	BCCA N=59	0.91 (0.59-1.39)	0.658	
	TCGA N=31	1.67 (1.06-2.64)	0.027	

Figure 48 Survival and Cohort

Kaplan-Meier survival curve for time after diagnosis (**a**) for four each of the sub-cohorts; SHGSOC (blue), BCCA (red), TCGA (purple), and AOCS (green). (b) A Cox proportional hazards model comparing the survival of the sub-cohorts adjusting for age, stage at diagnosis, HRD status, and WGD**;** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing (Benjamini and Hochberg 1995) correction was performed on all p values.

Survival Impact of the Most Disrupted Gene

In chapter 4, LSAMP was shown to be the most SV disrupted gene. Lower levels of LSAMP expression have been reported to be linked to shorter survival times in osteosarcoma patients and a study reported that its re-expression can inhibit tumour development in cell line experiments (Barøy et al. 2014; Kresse et al. 2009). To investigate if disruption of LSAMP impacts patient survival in HGSOC, samples with SV disrupted LSAMP were compared to samples without disrupted LSAMP. Samples with disrupted LSAMP appear at first sight to have shorter survival times after diagnosis (Figure 49a). However when modelled the effect on survival time is modest and does not reach statistical significance once adjusted for age, stage at diagnosis, HRD status, and WGD (Figure 49b). This suggests that the difference observed in the Kaplan-Meier plot can be explained by the covariates included in the Model.



b

Condition		Hazard ratio	P value	Forest plot
Age	N=279	1.01 (1.00-1.03)	0.09	
Stage	N=279	1.56 (1.19-2.05)	0.001	
HRD	Absent N=118	Reference		
	Present N=161	0.47 (0.34-0.65)	0.001	
Cohort	SHGSOC N=110	Reference		
	AOCS N=79	1.82 (1.23-2.70)	0.003	
	BCCA N=59	0.97 (0.63-1.52)	0.907	
	TCGA N=31	1.70 (1.07-2.69)	0.023	
LSAMP	Non - disrupted N=160	Reference		
	Disrupted N=119	1.17 (0.85-1.61)	0.325	

Figure 49 Expression of LSAMP and survival

Kaplan-Meier survival curve for time after diagnosis (**a**) for samples with (Blue) and without (Red) the LSAMP gene disrupted by an SV. (b) A Cox proportional hazards model

comparing samples with and without SV disrupted LSAMP adjusting for age, stage at diagnosis, HRD status, and WGD; a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values.

Amplification and Depletion of Chromosome Arms

Since many cSV events are large, sometimes encompassing entire chromosomes, it is of interest to explore the survival effects associated with other large-scale somatic alterations. Work conducted by Ailith Ewing utilising GISTIC (Mermel et al. 2011) shows there was significant recurrent amplification or depletion affecting 29 chromosome arms. In Chapter 3, chromosome 19 was found to be enriched for multiple cSV types and was also the only chromosome with both arms significantly amplified and depleted in the GISTIC data. The survival impacts of all chromosome arms showing significant amplification or depletion was tested, but none of these recurrent events had a significant effect on survival time (Figure 77 in appendix).

Gene Expression and Survival Impacts of Whole Genome Duplication

Another common large-scale somatic alteration in tumours is WGD, and in chapter 3 WGD was found to be present in nearly 50% of the combined HGSOC cohort. It has been reported that WGD happens early in tumour development and influences the subsequent development of the tumour (Quinton et al. 2021; Lens and Medema 2018; Fujiwara et al. 2005). WGD has also been reported to be associated with worse overall survival in a pan-cancer study of 9,692 advanced cancer patients diagnosed with one of 55 cancer types (Bielski et al. 2018). To investigate if HGSOC samples with WGD have worse overall survival and altered gene expression, samples with and without WGD were compared.

Figure 50a shows that among all protein-coding genes, only 95 genes exhibited a significant increase in expression in the presence of WGD, while 14 genes showed a

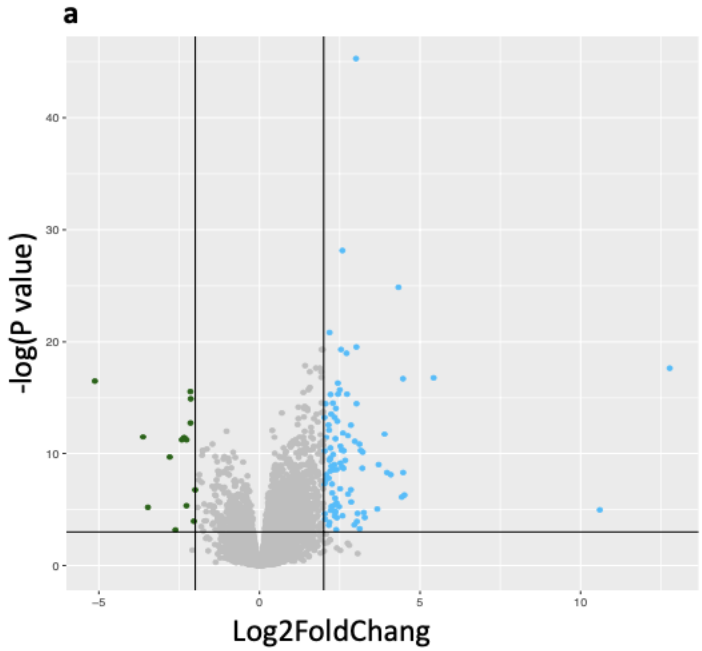
decrease in expression. The genes with the largest increase in expression are testis specific and olfactory genes, and these gene families are commonly observed to have increased expression in cancer gene expression studies (Jay et al. 2021; Fratta et al. 2011; Gjerstorff and Ditzel 2008; Weber et al. 2018; X. Zhang et al. 2007). Two KEGG gene pathways (M. Kanehisa and Goto 2000), the gastric cancer and melanoma pathways, were found to be marginally enriched in the genes with increased gene expression in the present of WGD.

The base mean shown in Figure 50b is the average of normalized counts of that gene's expression across all samples in the combined cohort with RNA expression data. A smaller base mean shows that there is lower expression across all the samples, and means that small changes in gene expression can be significant. Given that the base mean is less than 500 for all of the genes with the largest log fold change, it is likely that even modest absolute changes in gene expression between samples could lead to significant differences. This is also true for the genes showing the largest decreases in expression, so it is important to examine the base mean values as well as the magnitude and significance of expression changes.

One of the genes with the most significant change in expression (Figure 50d) was TPPP3 which shows a modest increase in expression in samples with WGD. This gene has been suggested to be a biomarker of HGSOC as its expression is highly enriched in fallopian tube tissue, and can be used to distinguish HSCOC from other cancers types such as ovarian clear cell carcinoma (A. J. Shih et al. 2018; Acland et al. 2020). The HP gene, which is highly expressed across the combined cohort samples with a base mean of over 9000, was found to be among the top 10 genes with the largest decrease in gene expression in samples with WGD. HP is normally expressed in liver and adipose tissue (Thul and Lindskog 2018; Noguchi et al. 2017; Lonsdale et al. 2013), but elevated HP is reported to be linked to poor prognosis in some solid cancer types (Tai et al. 2017). However, when comparing the survival of samples with and without WGD the presence of WGD had little impact on overall survival (Figure 50).

Given that WGD does significantly alter the expression of some genes, when we assess the impact of cSV types on gene expression we should also account for the presence or absence of WGD in samples, as a covariate in the differential expression analysis.

a



b

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GAGE12C	41.97	12.77	1.89	6.75	$2.16x10^{-08}$
TSPY3	5.04	10.60	2.92	3.63	$6.97x10^{-03}$
GAGE1	244.42	5.42	0.82	6.58	$5.10x10^{-08}$
OR51A2	9.75	4.52	1.10	4.09	$1.85x10^{-03}$
CT45A7	35.43	4.47	0.95	4.68	$2.49x10^{-04}$
GAGE2A	461.70	4.47	0.68	6.56	$5.52x10^{-08}$
MAGEA9	23.95	4.43	1.10	4.02	2.23x10-03
CTAG1B	154.43	4.33	0.55	7.91	1.59x10-11
CTAG1A	328.61	4.09	0.88	4.63	3.01x10-04
POU3F3	93.61	3.97	0.85	4.68	2.49x10-04

c

Largest decrease in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
AL358075.4	119.89	-5.13	0.79	-6.52	$6.81X10^{-08}$
CHRNB3	9.09	-3.62	0.66	-5.51	$1.02X10^{-05}$
CSN2	10.38	-3.47	0.94	-3.71	$5.56X10^{-03}$
ANKRD30A	19.84	-2.79	0.55	-5.05	$6.24X10^{-05}$
OR51A4	23.52	-2.62	0.92	-2.84	$4.19X10^{-02}$
FABP4	586.99	-2.42	0.44	-5.44	$1.34X10^{-05}$
GLYATL2	112.94	-2.35	0.43	-5.49	$1.06X10^{-05}$
ZFP91-CNTF	115.84	-2.27	0.60	-3.76	$4.81X10^{-03}$
OPCML	80.04	-2.27	0.42	-5.44	$1.34X10^{-05}$
HP	9367.90	-2.15	0.34	-6.35	$1.73X10^{-07}$

d

Most Significant

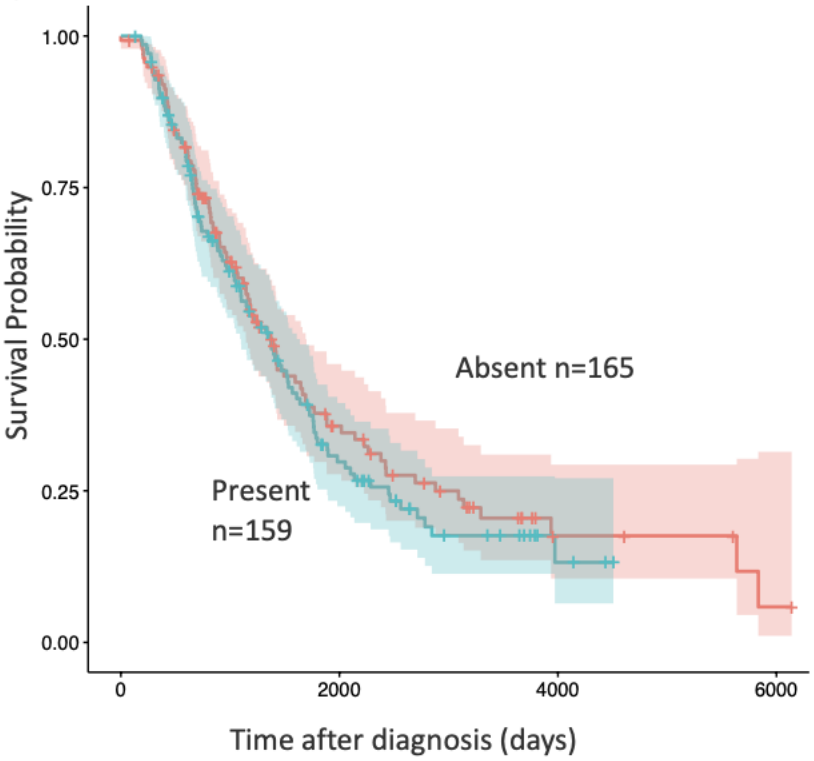
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
CRLF1	362.75	3.01	0.29	10.25	$2.17X10^{-20}$
TNNT3	263.81	2.58	0.31	8.36	$5.95X10^{-13}$
CTAG1B	154.43	4.33	0.55	7.91	$1.59X10^{-11}$
TMEM59L	139.89	2.18	0.30	7.36	$8.94X10^{-10}$
GCGR	24.08	3.02	0.42	7.15	$3.25X10^{-09}$
RAPSN	64.84	1.93	0.27	7.09	$4.07X10^{-09}$
DNER	84.41	1.97	0.28	7.06	$4.07X10^{-09}$
CYP2W1	132.34	2.54	0.36	7.06	$4.07X10^{-09}$
HRH3	11.56	2.71	0.39	6.99	$5.68X10^{-09}$
TPPP3	1466.01	1.42	0.21	6.82	$1.72X10^{-08}$

e

Gene Pathway enrichment (Increased expression)

Pathway	p.adjust	Count	GroupSize
Gastric cancer	0.015	4	149
Melanoma	0.015	3	72
Nicotine addiction	0.059	2	40
Breast cancer	0.059	3	147
PI3K-Akt signaling pathway	0.080	4	354
Neuroactive ligand-receptor interaction	0.080	4	367
Rap1 signaling pathway	0.080	3	210
Chemical carcin–genesis - receptor activation	0.080	3	212
Regulation of actin cytoskeleton	0.081	3	229
Ras signaling pathway	0.081	3	236

f



g

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.07	
Stage	N=280	1.54 (1.17-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.79 (1.22-2.62)	0.003	
	BCCA N=59	0.91 (0.59-1.39)	0.658	
	TCGA N=31	1.67 (1.06-2.64)	0.027	
WGD	Absent N=140	Reference		
	Present N=140	0.93 (0.69-1.25)	0.627	

Figure

50 Impact of WGD on gene expression and survival

The change in gene expression between the samples with and without whole genome duplication (WGD) corrected for sub-cohort. The X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with WGD (blue dots). Negative values represent a decrease in gene expression in samples with WGD (green dots). There were 95 genes with significantly increased expression and 14 genes with decreased expression. Expression data was used from 110 samples with WGD and 94 samples without WGD. The Y axis represents the -log10 p-values. The horizontal line represents the threshold for significance (0.05) with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are grey. Changes in gene expression were assessed using the DESeq2 package (Love, Huber, and Anders 2014). (b) Summary statistics for the ten genes with the greatest increase in gene expression. (c) Summary statistics for the ten genes with the greatest decrease in gene expression. (d) Summary statistics for the ten genes with the most significant changes in gene expIsion. (e) Gene set enrichment of genes with increased expression in samples with WGD using the KEGG terms (M. Kanehisa and Goto 2000; Minoru Kanehisa 2019; Minoru Kanehisa et al. 2023). (f) Kaplan-Meier survival curve for time after diagnosis in days for samples with WGD (blue) and without (red). (g) A Cox proportional hazards model comparing the survival of the samples with WGD to samples without WGD adjusting for age, stage at diagnosis, HRD status and sub-cohort; a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values.

Gene Expression and Survival Impact of Chromothripsis

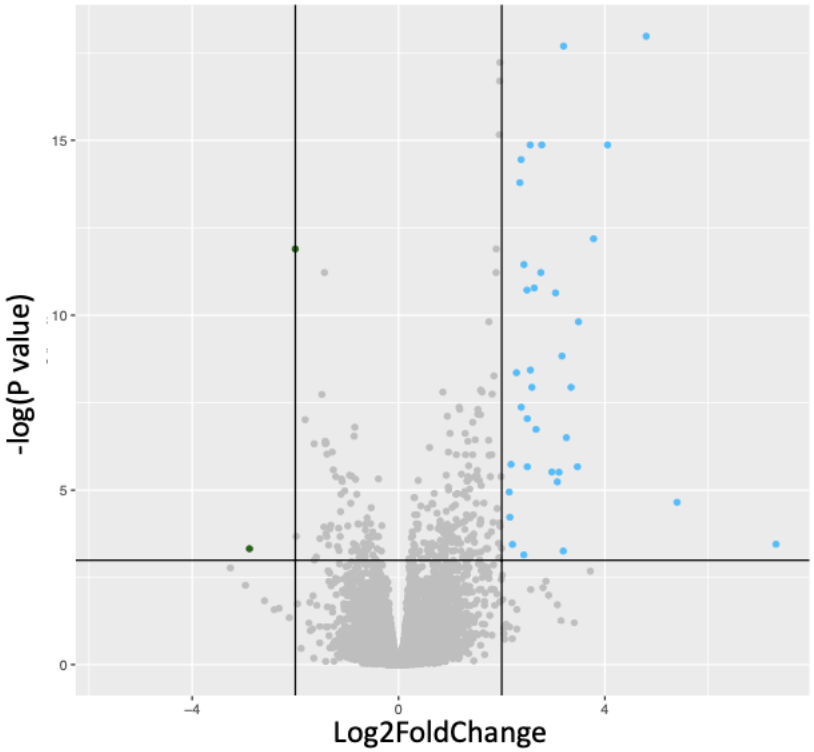
Chromothripsis has been reported to be associated with shorter overall survival in myeloid leukemia, multiple myeloma, and neuroblastoma (Fontana et al. 2018; Forment, Kaidi, and Jackson 2012; Kloosterman, Koster, and Molenaar 2014; Magrangeas et al. 2011; Molenaar et al. 2012). In contrast, in metastatic colorectal cancer it is reported that there is no impact on overall survival time, but chromothripsis is associated with shorter

progression-free survival (Skuja et al. 2017). Here, I investigate the impacts of chromothripsis on overall survival, progression-free survival and gene expression in HGSOC.

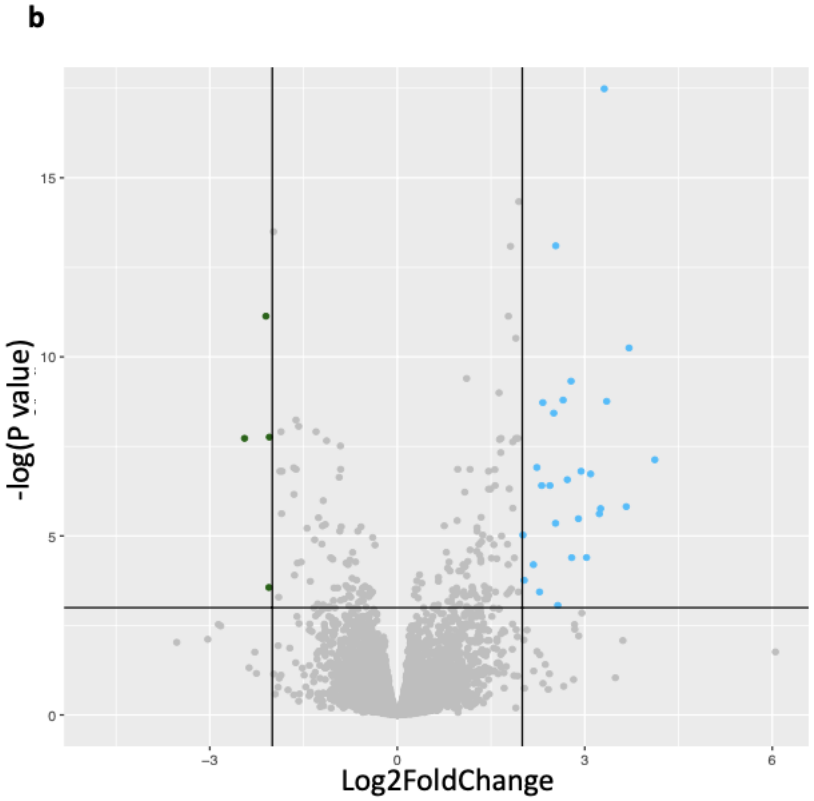
Chromothripsis had very little consistent effect on gene expression. In samples with chromothripsis, only 74 genes showed significantly increased expression and 35 genes showed a decrease in expression after accounting for WGD (Figure 51b). No KEGG pathways were found to be significantly enriched in genes with increased or decreased gene expression. One of the most significant gene expression changes (Figure 51e) was seen for SLC26A2 which showed a modest increase in expression (log2FC=1.11) in samples with chromothripsis. SLC26A2 expression is regulated in part by PAX8 a transcription factor frequently overexpressed in HGSOC (Adler et al. 2017; Cheung et al. 2011; Di Palma et al. 2014). Furthermore, *in vitro* and *in vivo* experiments with knockdown of PAX8 subsequently decreased expression of SLC26A2 and impaired tumorigenesis (Adler et al. 2017; Cheung et al. 2011; Di Palma et al. 2014).

The impact of chromothripsis on survival was not significant (Figure 51f and 51g). Additionally the impact of chromothripsis on progression free survival in the combined cohort was also not significant. (Progression free survival curves for all cSV types were non-significant and can be found in Figure 72 appendix.)

a


-log(P value)
15
10
5
0
−4
0
4
Log2FoldChange

b


-log(P value)
15
10
5
0
−3
0
3
6
Log2FoldChange

c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GAGE12C	32.86	6.05	2.20	2.75	1.72E-01
CTAG2	141.63	4.12	0.85	4.83	8.03E-04
CCL15	4.94	3.71	0.66	5.64	3.54E-05
XAGE1A	130.98	3.66	0.82	4.44	2.97E-03
TRIM51GP	6.76	3.61	1.22	2.96	1.25E-01
GDF1	39.68	3.49	1.62	2.15	3.54E-01
CTAG1B	265.25	3.35	0.63	5.29	1.57E-04
LRRC31	22.63	3.31	0.47	7.09	2.55E-08
GAGE2A	908.17	3.25	0.74	4.42	3.14E-03
GAGE12F	2621.25	3.23	0.74	4.38	3.64E-03

d

Largest decrease in expression

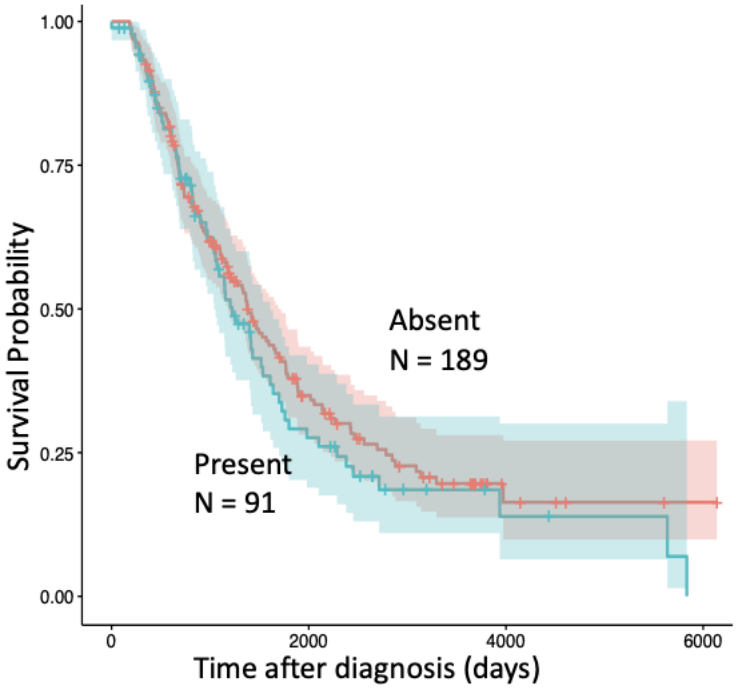
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
MBD3L2	8.88	-3.53	1.21	-2.92	1.32E-01
TRIM49D1	35.88	-3.03	1.02	-2.99	1.21E-01
POU3F3	71.05	-2.86	0.89	-3.22	7.99E-02
AC090227.1	10.30	-2.83	0.89	-3.19	8.33E-02
WIF1	86.82	-2.44	0.49	-4.99	4.42E-04
TRIM49C	14.96	-2.37	0.98	-2.42	2.68E-01
NPFF	5.06	-2.28	0.83	-2.75	1.73E-01
AC073612.1	7.95	-2.25	0.99	-2.28	3.13E-01
HOXD11	22.27	-2.10	0.36	-5.84	1.46E-05
AGTR2	13.45	-2.05	0.55	-3.71	2.84E-02

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
LRRC31	22.63	3.31	0.47	7.09	2.55E-08
SLC5A9	35.56	1.95	0.30	6.53	5.95E-07
NUP210L	47.06	-1.98	0.31	-6.34	1.38E-06
SPINK1	40.93	2.54	0.41	6.24	2.04E-06
F5	292.16	1.81	0.29	6.20	2.07E-06
HOXD11	22.27	-2.10	0.36	-5.84	1.46E-05
GREM2	119.69	1.78	0.30	5.83	1.46E-05
PNLDC1	55.09	1.90	0.33	5.70	2.70E-05
CCL15	4.94	3.71	0.66	5.64	3.54E-05
SLC26A2	3933.04	1.11	0.20	5.47	8.31E-05

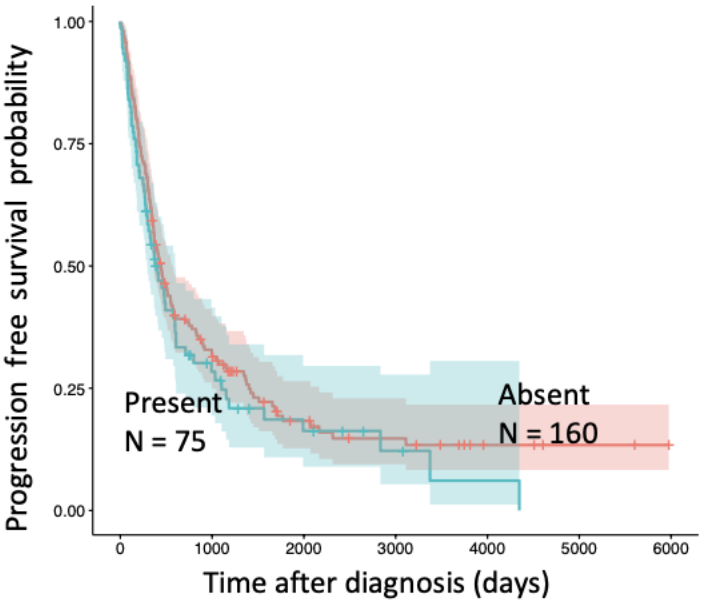
f



g

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.063	
Stage	N=280	1.54 (1.17-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.94 (0.70-1.27)	0.684	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.81 (1.23-2.67)	0.003	
	BCCA N=59	0.89 (0.57-1.37)	0.592	
	TCGA N=31	1.66 (1.05-2.62)	0.031	
Chromothripsis	Absent N=189	Reference		
	Present N=91	0.92 (0.66-1.28)	0.613	

h



i

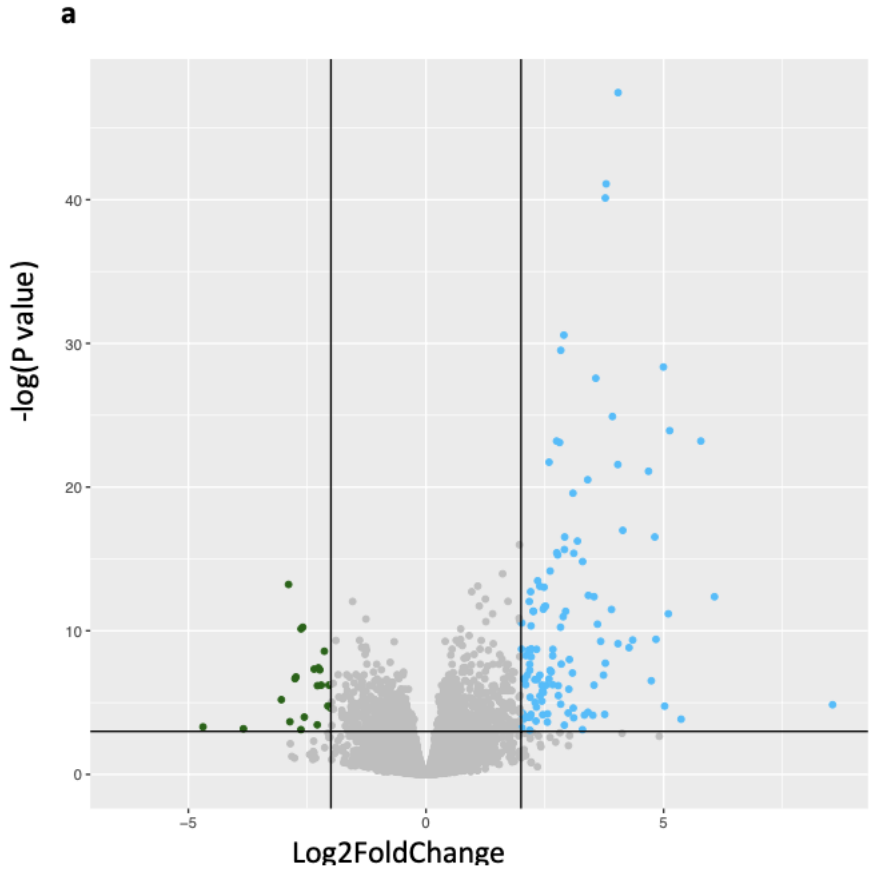
Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99-1.02)	0.421	
Stage	N=235	1.34 (1.0.6-1.71)	0.015	
HRD	Absent N=90	Reference		
	Present N=145	1.24 (0.42-0.81)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.24 (0.93-1.67)	0.145	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.20 (2.19-4.68)	0.001	
	BCCA N=58	1.40 (0.94-2.09)	0.097	
	TCGA N=0			
Chromothripsis	Absent N=160	Reference		
	Present N=75	0.96 (0.69-1.34)	0.816	

Figure 51 Impact on gene expression and survival of chromothripsis

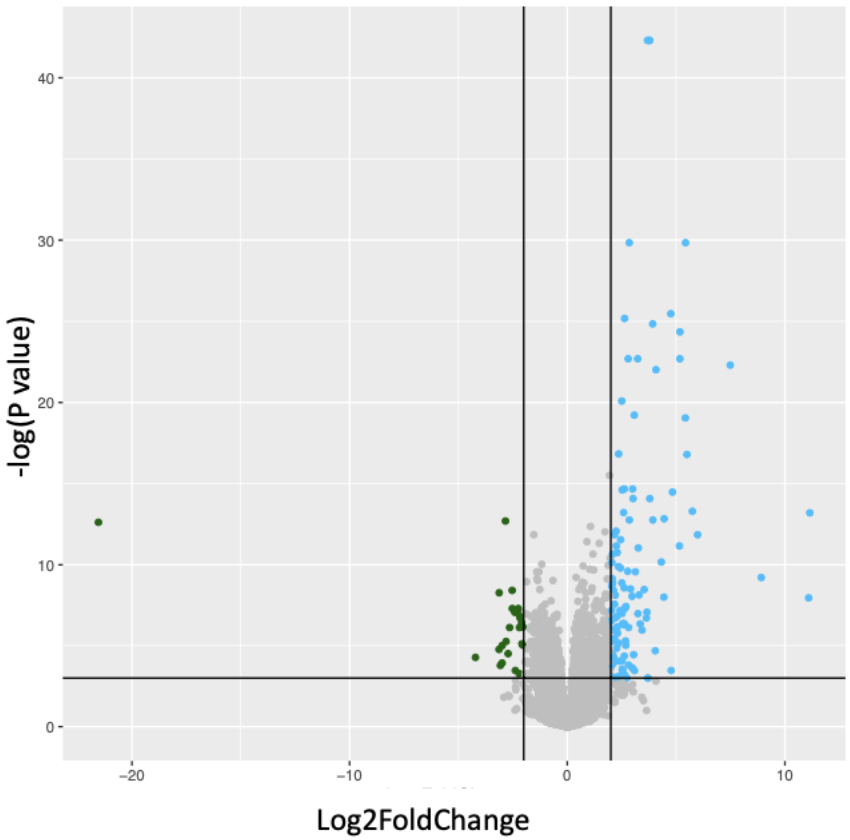
The change in gene expression between the samples with and without chromothripsis corrected for sub-cohort. There were 38 genes with increased expression and 3 genes with decreased expression (**a**)**.** The change in gene expression between the samples with and without chromothripsis corrected for sub-cohort and WGD (**b**) .There were 74 genes with increased expression and 35 genes with decreased expression. Expression data was used from 74 samples with chromothripsis and 130 samples without chromothripsis. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with chromothripsis (blue dots). Negative values represent a decrease in gene expression in samples with chromothripsis (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are shown in grey. Changes in gene expression was assessed using the DESeq2 package (Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest increase in geneIpression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the most significant change in gI expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with chromothripsis (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with chromothripsis to samples without chromothripsis adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values. No KEGG pathways were found to be significantly ($p<0.05$) enriched in genes with increased or decreased gene expression. Kaplan-Meier progression free survival curve for time after diagnosis in days for samples with chromothripsis (blue) and without (red) **h**. A Cox proportional hazards model comparing the progression free survival the samples with chromothripsis to samples without chromothripsis adjusting for age, stage at diagnosis, HRD status and sub-cohort **i**

Gene Expression and Survival Impact of Rigma

When initially described rigma was reported to be associated with significantly worse survival (Hadi et al. 2020), but this is the only study of rigma and survival, so further investigation is needed. In HGSOC the effect of rigma on survival showed no clear effect (Figure 52f and 52g). There were 138 genes with significantly increased expression and 22 genes with decreased expression in the presence of rigma (Figure 52a), and once the effect of WGD had been accounted for there were 149 genes with increased expression and 28 genes with decreased expression (Figure 52b). No KEGG pathways were found to be significantly enriched in differentially Expressed genes.



b



c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GDF1	39.68	11.15	1.88	5.94	1.85x10-06
GAGE12C	681.73	11.09	2.33	4.77	3.48x10-04
GAGE12H	268.12	8.91	1.75	5.10	9.93x10-05
ISX	13.09	7.49	1.01	7.39	2.06x10-10
CDH9	18.65	5.99	1.06	5.66	7.11x10-06
CTAG2	141.63	5.75	0.96	5.97	1.68x10-06
GAGE2A	902.16	5.50	0.84	6.57	5.07x10-08
PAGE2	283.48	5.43	0.64	8.46	1.10x10-13
SFTA3	15.71	5.42	0.79	6.91	5.34x10-09
EPS8L3	44.76	5.18	0.67	7.71	2.67x10-11

d

Largest decrease in expression

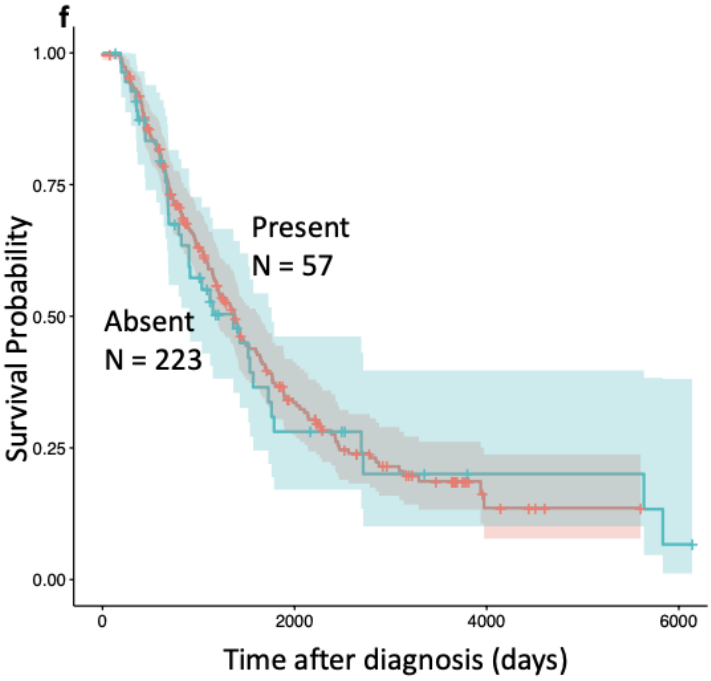
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
TSPY3	5.04	-21.54	3.70	-5.82	3.30x10-06
OR51A4	23.52	-4.22	1.20	-3.53	1.40x10-02
CHRNB3	9.09	-3.13	0.84	-3.73	8.57x10-03
PIP	15.41	-3.13	0.64	-4.85	2.54x10-04
SPRR1B	48.37	-3.07	0.93	-3.30	2.29x10-02
OR4N2	20.46	-2.99	0.79	-3.81	6.81x10-03
USP17L11	30.45	-2.99	0.89	-3.37	1.98x10-02
OR1S2	25.23	-2.92	1.32	-2.20	1.64x10-01
SOX2	203.85	-2.84	0.49	-5.83	3.06x10-06
PIK3C2G	70.95	-2.81	0.72	-3.92	5.24x10-03

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
COX8C	35.62	3.69	0.37	9.85	4.25x10-19
FAM151A	12.63	3.76	0.38	9.86	4.25x10-19
CLDN18	14.12	3.79	0.38	9.85	4.25x10-19
NUP210L	47.06	2.85	0.34	8.45	1.10x10-13
PAGE2	283.48	5.43	0.64	8.46	1.10x10-13
KCNQ2	119.13	4.76	0.60	7.90	8.70x10-12
F5	292.16	2.63	0.34	7.84	1.17x10-11
LHFPL4	64.05	3.92	0.50	7.78	1.62x10-11
EPS8L3	44.76	5.18	0.67	7.71	2.67x10-11

f



g

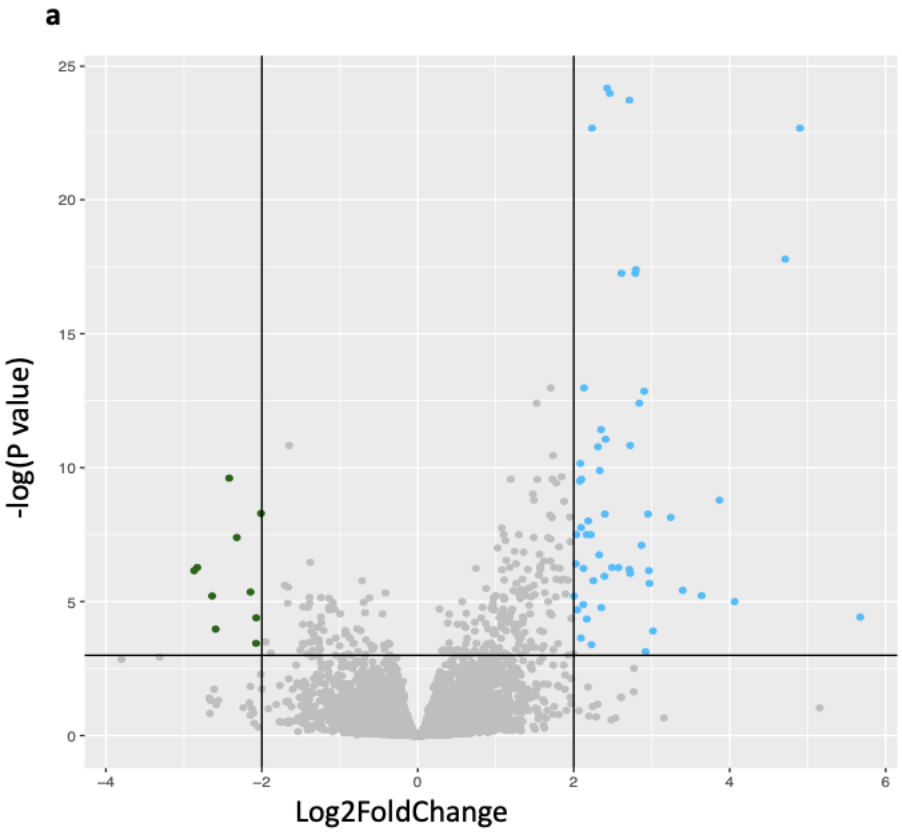
Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.079	
Stage	N=280	1.54 (1.18-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.93 (0.69-1.24)	0.614	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.80 (1.22-2.66)	0.003	
	BCCA N=59	0.92 (0.60-1.42)	0.710	
	TCGA N=31	1.71 (1.08-1.24)	0.023	
Rigma	Absent N=223	Reference		
	Present N=57	1.11 (0.76-1.63)	0.577	

Figure 52 Impact on gene expression and survival of rigma

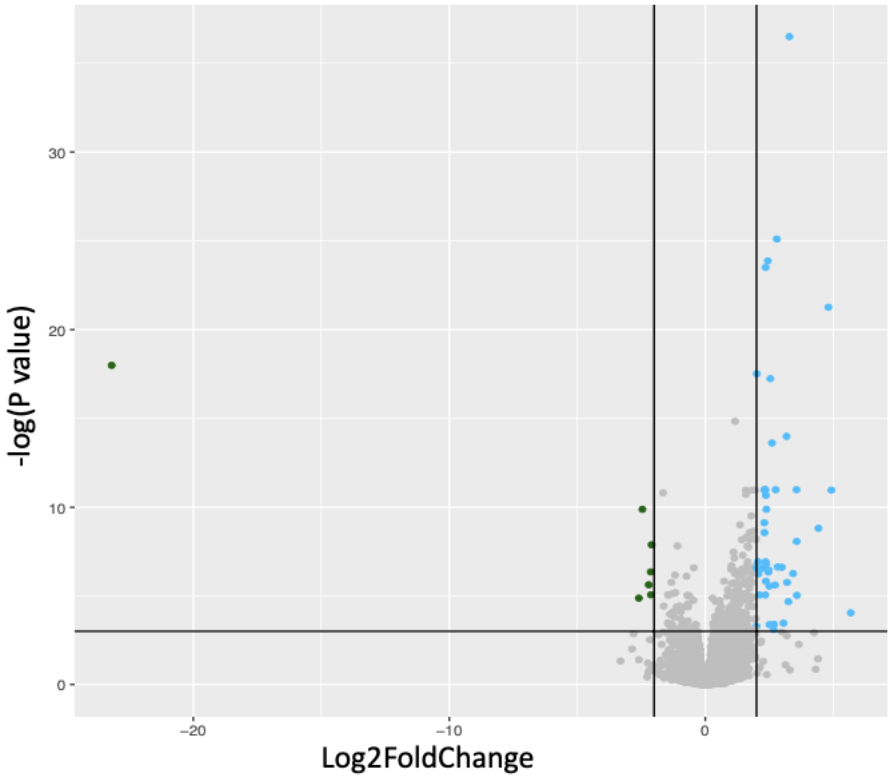
The change in gene expression between the samples with and without rigma corrected for sub-cohort. There were 62 genes with increased expression and 11 genes with decreased expression (**a**)**.** The change in gene expression between the samples with and without rigma corrected for sub-cohort and WGD (**b**) .There were 60 genes with increased expression and 7 genes with decreased expression. Expression data was used from 56 samples with rigma and 148 samples without rigma. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with rigma (blue dots). Negative values represent a decrease in gene expression in samples with rigma (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest increase iIene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the most significant changIn gene expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with rigma (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with rigma to samples without rigma adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was perfor170oron all p values. No kegg pathways were found to be significantly ($p<0.05$) enriched in genes with increased or decreased gene expression.

Gene Expression and Survival Impact of Pyrgo

As for rigma, when pyrgo was initially described it was reported to be associated with significantly worse survival (Hadi et al. 2020) but this has not been studied further. Once the effect of WGD had been accounted for, there were 60 genes with increased expression and 7 genes with decreased expression (Figure 53b). One of the most significant expression changes found was the increase in expression of MIEN1 (log2FC=1.16) which has previously been implicated in survival in ovarian cancer (Figure 53e) (Cámara-Quílez et al. 2020). However, in HGSOC the effect of pyrgo on survival was not significant once age, stage, HRD and cohort were adjusted for (Figure 53f and 53g).



b



c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GDF1	39.68	5.68	1.50	3.79	1.72E-02
XAGE1A	302.05	4.93	0.87	5.65	1.72E-05
ACTL6B	14.76	4.81	0.65	7.38	5.84E-10
KRT76	17.42	4.42	0.85	5.18	1.48E-04
GAGE12H	16.97	4.40	1.76	2.49	2.36E-01
GAGE12C	41.97	4.31	2.20	1.96	4.23E-01
OR8U1	6.37	4.25	1.28	3.32	5.39E-02
HTN3	4.34	3.65	1.22	2.99	1.05E-01
GAGE13	469.45	3.58	0.87	4.11	6.44E-03
GAGE12F	434.70	3.57	0.72	4.98	3.10E-04

d

Largest decrease in expression

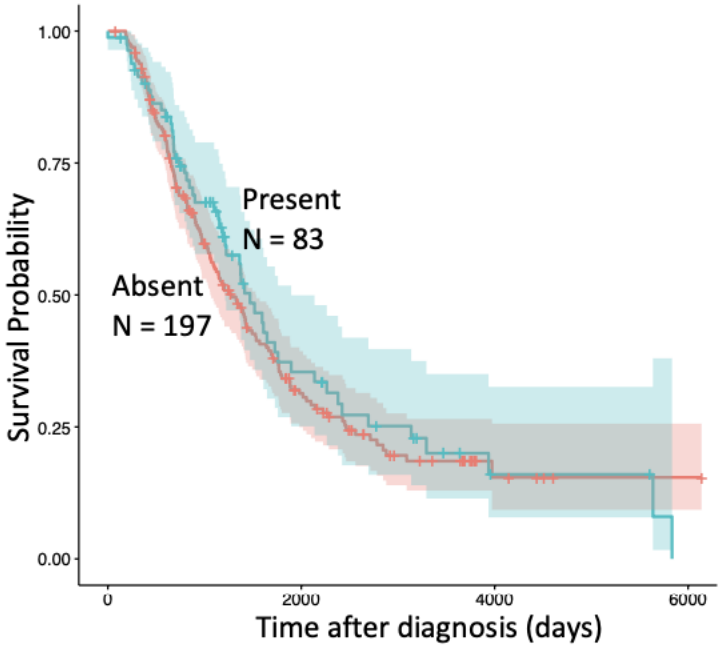
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
TSPY3	5.04	-23.19	3.36	-6.91	1.54E-08
TBC1D3K	15.57	-3.31	1.38	-2.40	2.67E-01
AC011455.2	33.73	-2.86	1.01	-2.83	1.36E-01
GLYATL3	11.90	-2.80	0.85	-3.29	5.72E-02
EPS8L3	44.76	-2.60	0.64	-4.05	7.53E-03
AC073612.1	7.95	-2.59	1.06	-2.45	2.49E-01
NPIPA8	1046.13	-2.46	0.45	-5.40	5.07E-05
SSX4B	85.92	-2.27	1.68	-1.36	6.53E-01
HOXC12	11.57	-2.24	0.97	-2.31	2.99E-01
MBD3L2	7.39	-2.23	1.26	-1.76	5.00E-01

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
PNMA5	43.75	3.28	0.35	9.37	1.42E-16
PNLDC1	55.09	2.79	0.35	7.99	1.25E-11
NUP210L	47.06	2.44	0.31	7.79	4.32E-11
CHRNB2	84.25	2.35	0.31	7.70	6.22E-11
ACTL6B	14.76	4.81	0.65	7.38	5.84E-10
TSPY3	5.04	-23.19	3.36	-6.91	1.54E-08
ACSBG1	69.55	2.01	0.30	6.82	2.47E-08
COX8C	35.62	2.54	0.38	6.76	3.24E-08
MIEN1	2382.03	1.16	0.18	6.39	3.59E-07
GDPD4	65.41	3.17	0.51	6.24	8.35E-07

f



g

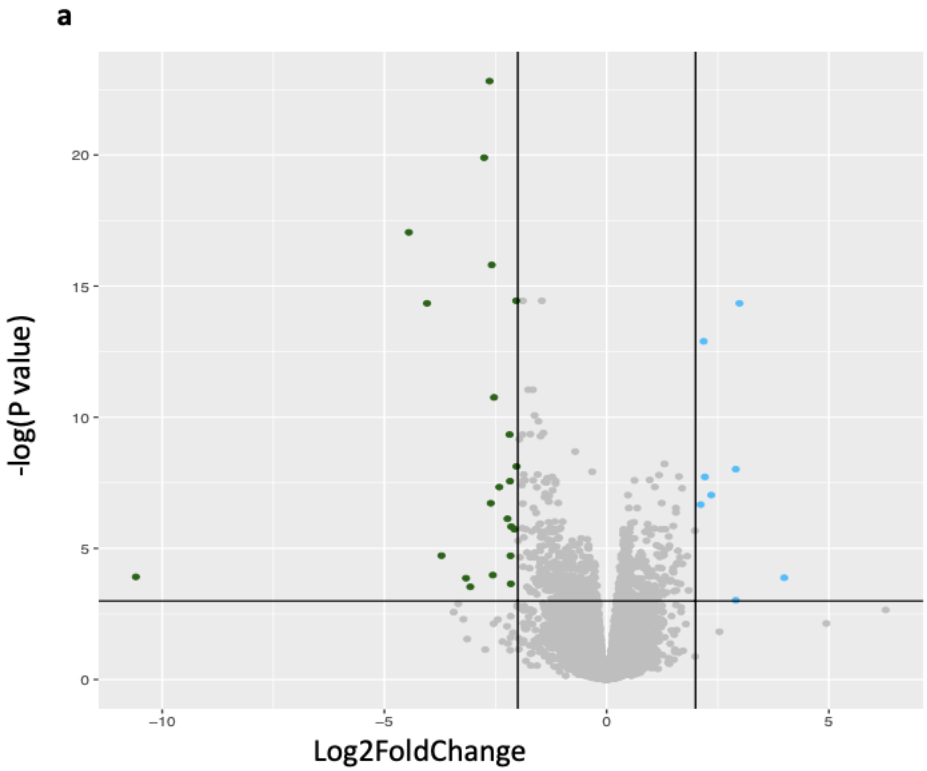
Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.063	
Stage	N=280	1.54 (1.17-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.91 (0.67-1.23)	0.538	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.78 (1.21-2.62)	0.003	
	BCCA N=59	0.91 (0.59-1.39)	0.592	
	TCGA N=31	1.65 (1.04-2.1)	0.031	
Pyrgo	Absent N=197	Reference		
	Present N=83	0.89 (0.64-1.25)	0.613	

Figure 53 Impact on gene expression and survival of pyrgo

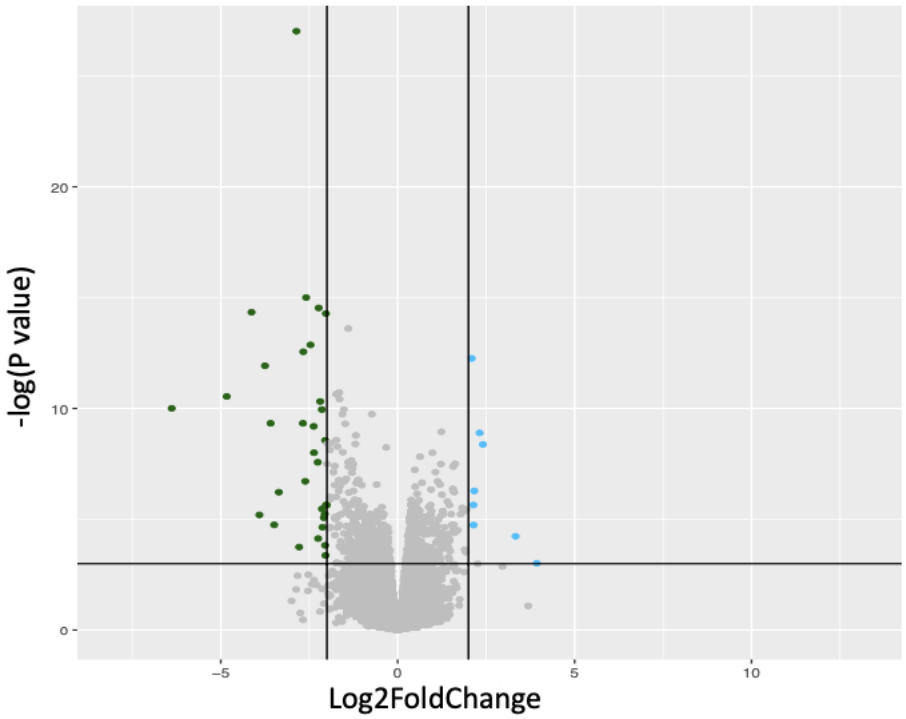
The change in gene expression between the samples with and without pyrgo corrected for sub-cohort. There were 62 genes with increased expression and 11 genes with decreased expression (**a**)**.** The change in gene expression between the samples with and without pyrgo corrected for sub-cohort and WGD (**b**) .There were 60 gene with increased expression and 7 genes with decreased expression. Expression data was used from 56 samples with pyrgo and 148 samples without pyrgo. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with pyrgo (blue dots). Negative values represent a decrease in gene expression in samples with pyrgo (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest incrIe in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the most significantIange in gene expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with pyrgo (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with pyrgo to samples without pyrgo adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was 175ororormed on all p values. No kegg pathways were found to be significantly ($p<0.05$) enriched in genes with increased or decreased gene expression.

Gene Expression and Survival Impact of Chromoplexy

In a study of 85 WGS myeloma samples, chromoplexy was found in 21 samples and was reported to increase the expression of MYC, E2F and G2M targets, and to reduce RAS signalling (Ashby et al. 2019). In the combined cohort, once WGD was accounted for, I observed that 11 genes had increased expression and 36 genes had decreased expression (Figure 54 b). Only two pathways were found to be marginally significantly enriched in genes with decreased expression (Figure 54f) and they have no strong connection to RAS signalling. In addition, there was no significant impact of chromoplexy on patient survival (Figure 54g and 54h).



b



c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
OR51A2	9.75	3.93	1.28	3.07	4.90E-02
OR11H1	5.14	3.69	2.15	1.72	3.35E-01
FOXR2	27.29	3.33	0.91	3.66	1.45E-02
ASIC5	2.28	2.96	0.99	2.99	5.65E-02
RNF17	29.88	2.41	0.48	5.07	2.30E-04
SPATA22	8.24	2.32	0.44	5.21	1.37E-04
SPRR1B	48.37	2.26	0.74	3.06	5.03E-02
SCGB2A2	127.01	2.16	0.48	4.48	1.87E-03
CT45A10	739.26	2.14	0.55	3.90	8.75E-03
KCNQ2	119.13	2.14	0.50	4.28	3.54E-03

d

Largest decrease in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
OR5D18	4.28	-6.39	1.16	-5.49	4.51E-05
APCS	8.49	-4.83	0.86	-5.62	2.64E-05
SMR3B	44.74	-4.13	0.64	-6.44	5.90E-07
TRIM49D1	35.88	-3.91	0.95	-4.09	5.53E-03
NKX2-1	21.36	-3.75	0.63	-5.91	6.62E-06
FABP7	18.57	-3.59	0.68	-5.32	8.86E-05
Z83844.1	58.99	-3.49	0.89	-3.90	8.67E-03
DCAF4L2	17.38	-3.36	0.75	-4.46	1.98E-03
GDF1	39.68	-3.00	1.56	-1.92	2.69E-01
ARGFX	3.54	-2.87	1.23	-2.34	1.59E-01

e

Most Significant

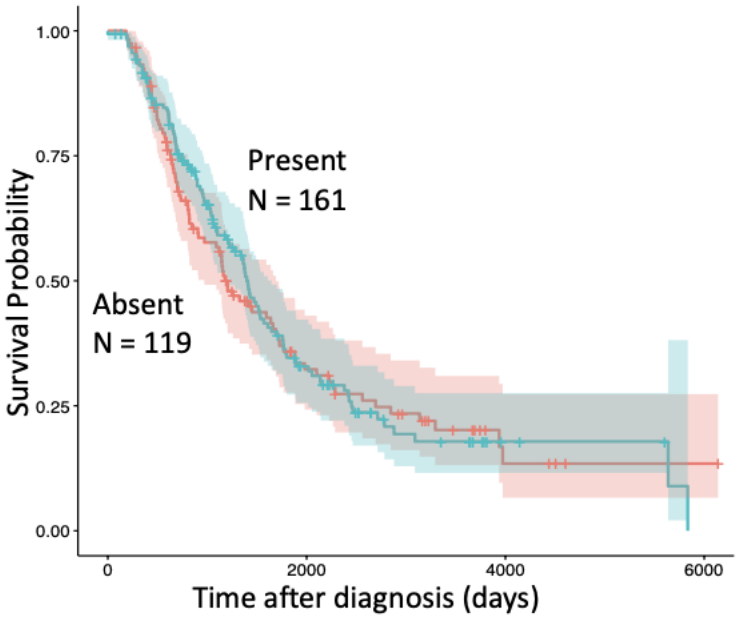
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
ATP1A2	1219.88	-2.86	0.34	-8.31	1.84E-12
OR7D4	23.65	-2.59	0.39	-6.64	3.05E-07
GRIA4	92.06	-2.23	0.34	-6.51	4.87E-07
SMR3B	44.74	-4.13	0.64	-6.44	5.90E-07
FAM133A	86.89	-2.03	0.32	-6.39	6.28E-07
ALS2CR12	135.17	-1.40	0.22	-6.26	1.23E-06
GCGR	24.08	-2.46	0.40	-6.12	2.58E-06
OLFM4	74.19	-2.67	0.44	-6.05	3.52E-06
NTS	329.94	2.09	0.35	5.98	4.75E-06
NKX2-1	21.36	-3.75	0.63	-5.91	6.62E-06

f

Gene Pathway enrichment (Decreased expression)

Pathway	P value	Count	GroupSize
Bile secretion	0.049	2	89
Neuroactive ligand-receptor interaction	0.049	3	367
Proximal tubule bicarbonate reclamation	0.121	1	23
Ascorbate and aldarate metabolism	0.121	1	30
Pentose and glucuronate interconversions	0.121	1	35
Aldosterone-regulated sodium reabsorption	0.121	1	37
Porphyrin metabolism	0.121	1	43
Carbohydrate digestion and absorption	0.121	1	47
Endocrine and other factor-regulated calcium reabsorption	0.121	1	53
Mineral absorption	0.121	1	60

g



h

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.063	
Stage	N=280	1.54 (1.17-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.45 (0.33-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.92 (0.68-1.24)	0.584	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.80 (1.22-2.66)	0.003	
	BCCA N=59	0.92 (0.60-1.41)	0.694	
	TCGA N=31	1.69 (1.07-2.69)	0.025	
Chromoplexy	Absent N=119	Reference		
	Present N=161	1.06 (0.78-1.45)	0.718	

Figure 54 Gene expression and survival chromoplexy

The changes in gene expression between the samples with and without chromoplexy corrected for sub-cohort. There were 8 genes with increased expression and 23 genes with decreased expression (**a**)**.** The changes in gene expression between the samples with and without chromoplexy corrected for sub-cohort and WGD (**b**) .There were 11 genes with increased expression and 36 genes with decreased expression. Expression data was used from 117samples with chromoplexy and 87 samples without chromoplexy. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with chromoplexy (blue dots). Negative values represent a decrease in gene expression in samples with chromoplexy (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest increase in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the most significI change in gene expression (**e)**. Gene set enrichment of genes with decreased expression in samples with chromoplexy using the KEGG terms **f** (M. Kanehisa and Goto 2000; Minoru Kanehisa 2019; Minoru Kanehisa et al. 2023). Kaplan-Meier survival curve for time after diagnosis in days for samples with chromoplexy (blue) and without (red) **g**. A Cox proportional hazards model comparing the survival of the samples with chromoplexy to samples without chromoplexy adjusting for age, stage at diagnosis, HRD status and sub-cohort **h** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values. No kegg pathways were found to be significantly ($p<0.05$) enriched in genes with increased or decreased gene expression.

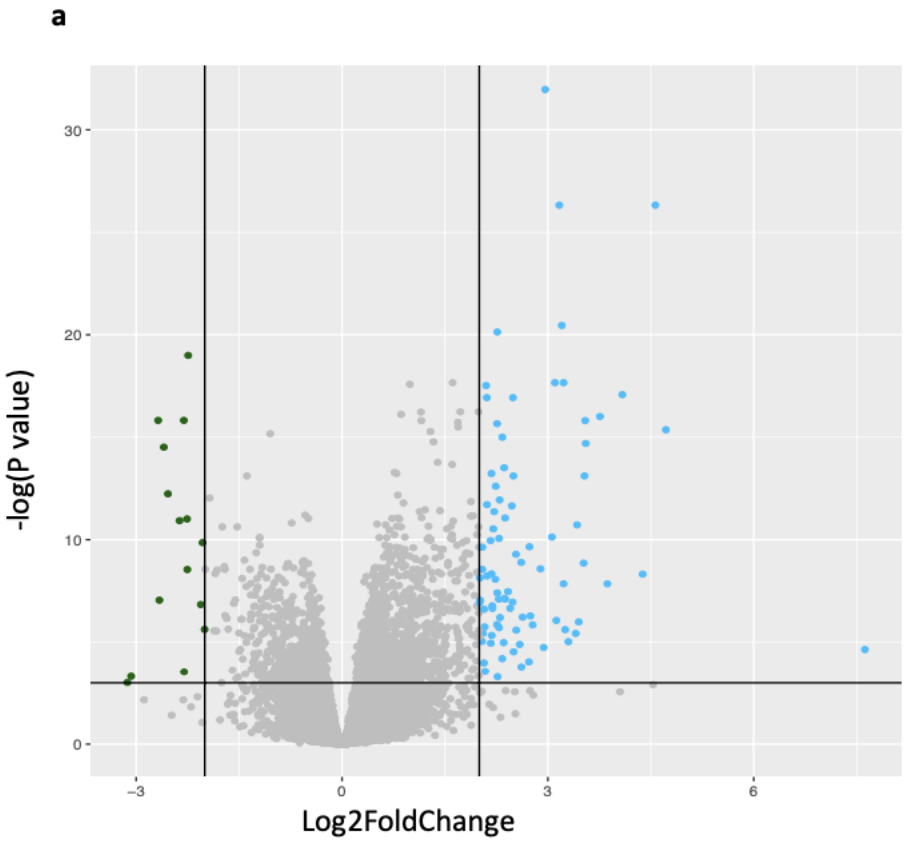
Gene Expression and Survival Impacts of Breakage Fusion Bridge Events

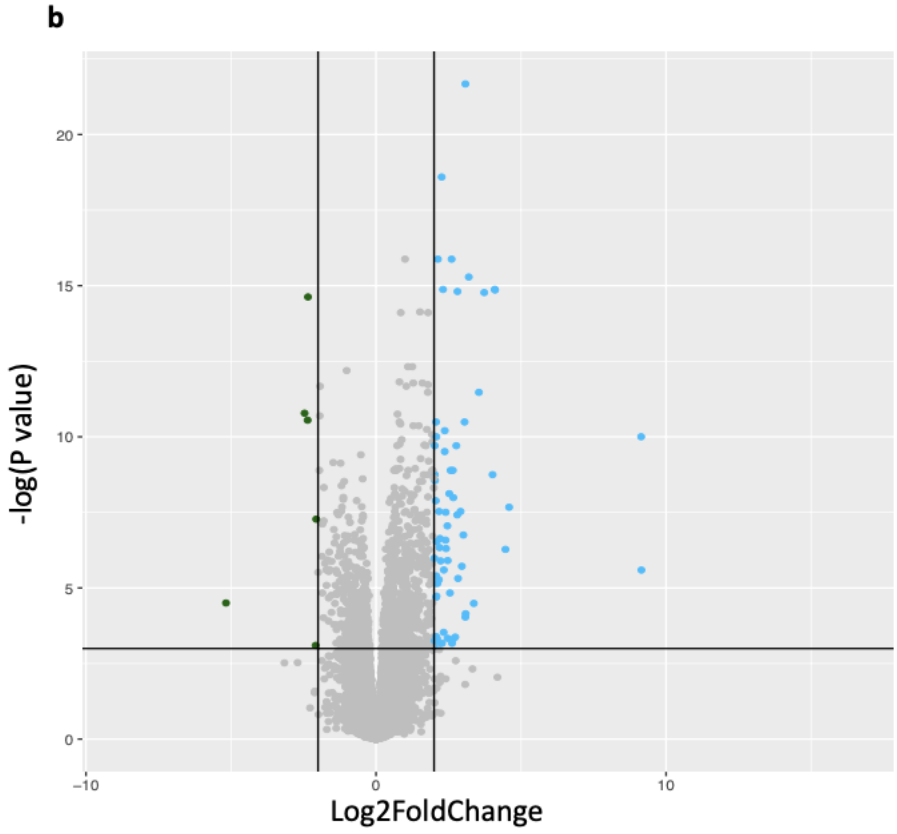
In Chapter 3, it was shown that BFB events were common and found in 76 HGSOC samples. While in Chapter 4 it was shown that BFB overlapped the known oncogene

CCNE1 in 24 of those samples. To investigate the impact of BFB on gene expression and overall survival, samples with and without BFB were compared. After adjusting for WGD, 96 genes were found with significantly increased expression and 7 genes with decreased expression. Interestingly, despite being a cSV involving amplifications of gene copy numbers, the overall effect of BFB on CCNE1 showed only a modest increase in expression (log2FC=0.77, pvalue $3.14x10^{-3}$) across all BFB samples, and CCNE1 does not appear in the most significantly altered genes (Figure 55).

The TMEM100 gene showed a large decrease in expression (Figure 55d) in samples with BFB, and TMEM100 is another gene regulated by the PAX8 transcription factor (Elias et al. 2016). Among the most significant changes in gene expression (Figure 55e) was a modest increase for the ERBB2 gene, which is reported to have amplified expression in 3-10% of ovarian cancer but was not reported to impact overall survival (Thouvenin et al. 2021). In samples with BFB, the gene PIGR showed a significant decrease in expression (log fold change -2.46, p value $2.08x10^{-5}$) and low expression of PIGR has been associated with worse survival outcomes in ovarian cancer and nasopharyngeal carcinoma (Qi, Li, and Sun 2016; Biswas et al. 2021). Despite these interesting changes in gene expression the overall impact of BFB on patient survival was not significant when modelled (Figure 55f).

To directly investigate the effect of BFB amplified CCNE1, samples with CCNE1 amplified by BFB were compared to the rest of the combined cohort, and the effect on survival was also found to be non-significant (Figure 55h and 55i).


a
30
20
10
0
-log(P value)
−3
0
3
6
Log2FoldChange


b
20
15
10
5
0
-log(P value)
−10
0
10
Log2FoldChange

c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GAGE12C	59.01	9.15	2.28	4.00	3.73E-03
GDF1	39.68	9.14	1.71	5.33	4.53E-05
FTHL17	5.93	4.59	0.98	4.69	4.66E-04
CDH9	18.65	4.46	1.05	4.25	1.88E-03
GAGE12H	9.71	4.19	1.75	2.39	1.29E-01
SLC2A2	27.08	4.10	0.64	6.39	3.53E-07
DDX53	28.26	4.09	0.64	6.43	3.46E-07
MAGEC2	97.24	4.02	0.80	4.99	1.59E-04
OR1N1	31.53	3.73	0.59	6.34	3.84E-07
EPS8L3	44.76	3.55	0.63	5.67	1.04E-05

d

Largest decrease in expression

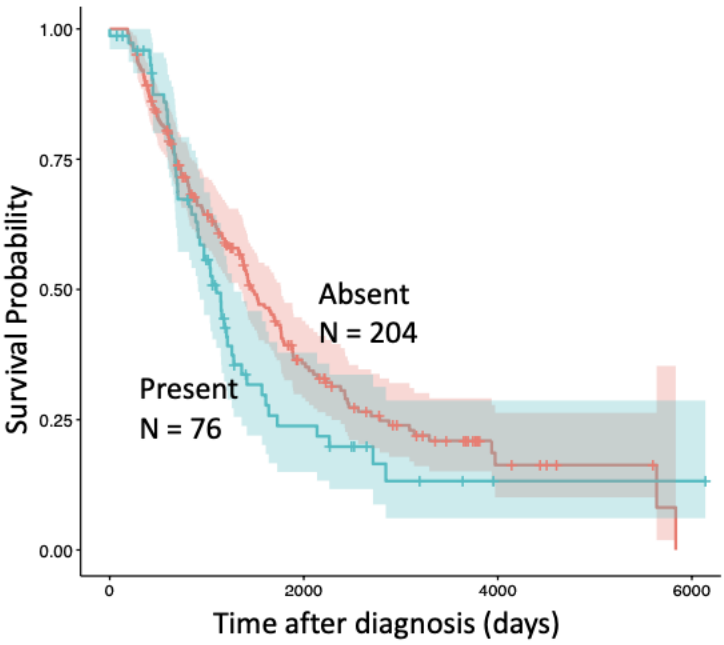
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
OR51A2	9.75	-5.17	1.43	-3.61	1.11E-02
OR1S2	25.23	-3.16	1.18	-2.68	8.04E-02
TAS2R16	22.17	-2.70	1.01	-2.68	7.97E-02
PIGR	2915.79	-2.46	0.44	-5.54	2.08E-05
SLC25A48	81.91	-2.36	0.43	-5.48	2.62E-05
TMEM100	1048.25	-2.35	0.37	-6.31	4.44E-07
GAGE12J	61.26	-2.27	1.40	-1.63	3.57E-01
FOXR2	27.29	-2.12	1.04	-2.04	2.16E-01
OR6N2	30.74	-2.11	1.02	-2.08	2.05E-01
SPINK8	2.83	-2.08	0.70	-2.99	4.50E-02

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
RBP2	21.64	3.08	0.40	7.65	3.86E-10
NUP210L	47.06	2.26	0.32	7.15	8.41E-09
ERBB2	11702.46	1.00	0.15	6.63	1.27E-07
CRLF1	443.34	2.13	0.32	6.64	1.27E-07
KRT13	195.38	2.61	0.39	6.70	1.27E-07
VRTN	17.56	3.20	0.49	6.51	2.30E-07
DDX53	28.26	4.09	0.64	6.43	3.46E-07
HSD17B2	44.95	2.31	0.36	6.41	3.47E-07
SLC2A2	27.08	4.10	0.64	6.39	3.53E-07
CLDN19	158.94	2.81	0.44	6.36	3.71E-07

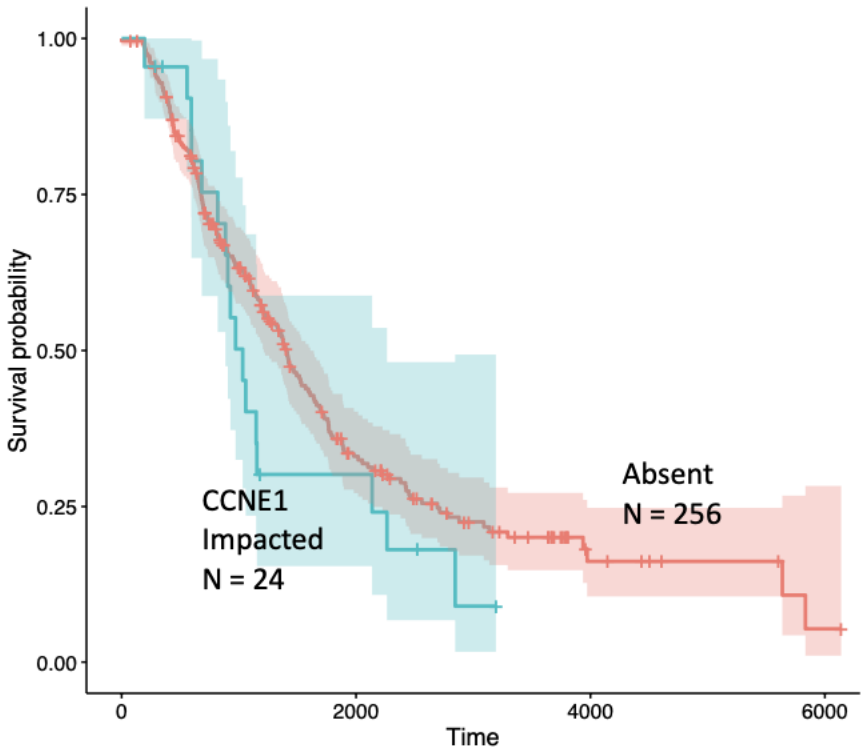
f



g

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.02 (1.00-1.03)	0.06	
Stage	N=280	1.54 (1.18-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.45 (0.33-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.94 (0.70-1.26)	0.668	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.79 (1.22-2.63)	0.003	
	BCCA N=59	0.90 (0.59-1.39)	0.694	
	TCGA N=31	1.65 (1.04-2.62)	0.025	
BFB	Absent N=119	Reference		
	Present N=161	0.92 (0.64 -1.32)	0.638	

h



i

Condition		Hazard ratio	P value	Forest plot
Age	N=279	1.01 (1.00-1.03)	0.063	
Stage	N=279	1.54 (1.18-2.02)	0.002	
HRD	Absent N=118	Reference		
	Present N=161	0.45 (0.33-0.63)	0.001	
Cohort	SHGSOC N=110	Reference		
	AOCS N=79	1.76 (1.19-2.58)	0.004	
	BCCA N=59	0.91 (0.60-1.40)	0.681	
	TCGA N=31	1.65 (1.04-2.61)	0.034	
CCNE1 BFB	Absent N=256	Reference		
	Disrupted N=24	0.90 (0.53-1.53)	0.695	

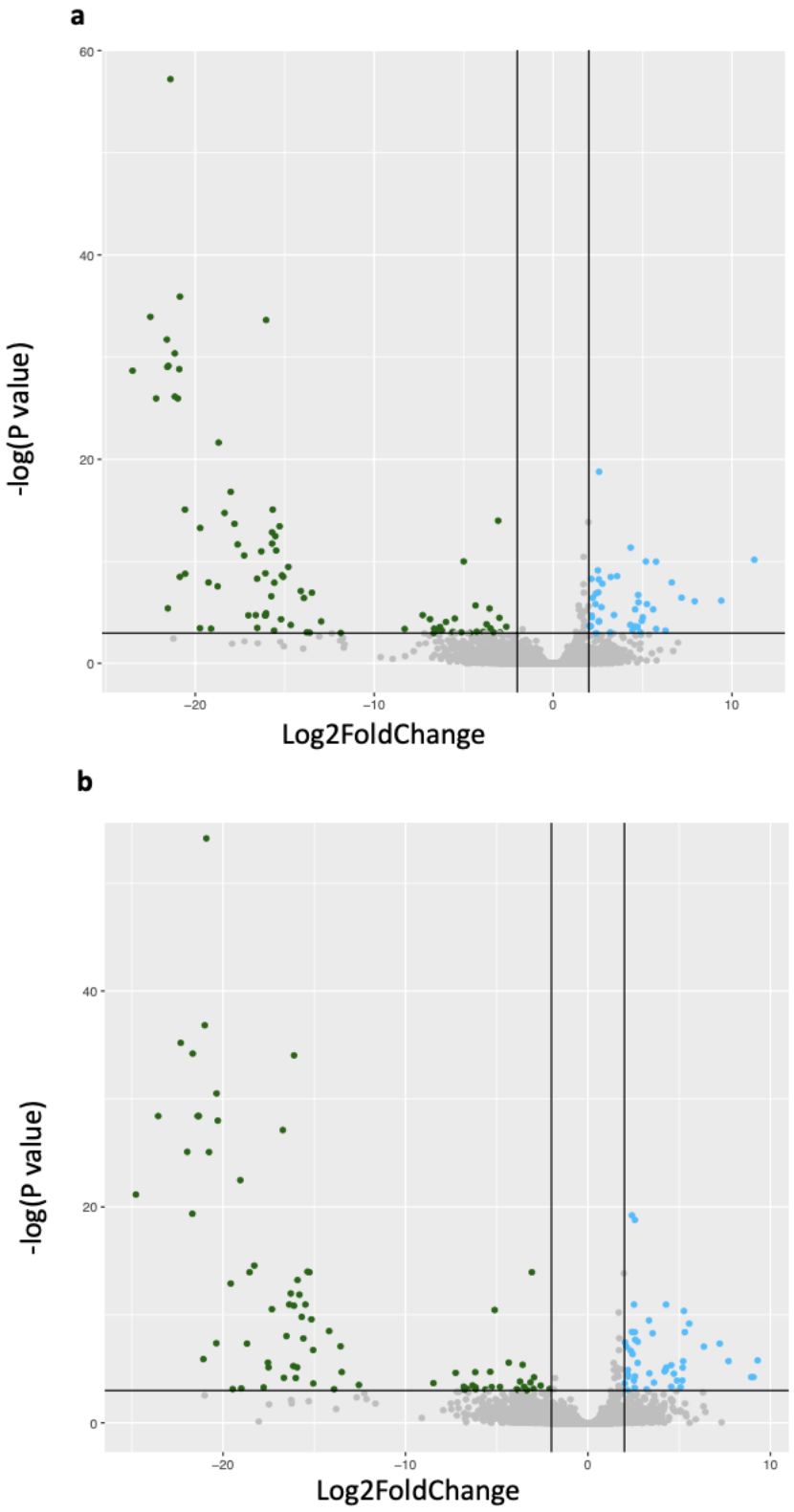
Figure 55 Gene expression and survival with BFB

The changes in gene expression between the samples with and without breakage fusion

bridges (BFB) corrected for sub-cohort. There were 90 genes with increased expression and 16 genes with decreased expression (**a**)**.** The changes in gene expression between the samples with and without BFB corrected for sub-cohort and WGD (**b**) .There were 96 genes with increased expression and 7 genes with decreased expression. Expression data was used from 59 samples with BFB and145 samples without BFB. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with BFB (blue dots). Negative values represent a decrease in gene expression in samples with BFB (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greaIt increase in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the most siIficant change in gene expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with BFB (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with BFB to samples without BFB adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Kaplan-Meier survival curve for time after diagnosis in days for samples with BFB disrupting CCNE1 (blue) and without BFB disrupting CCNE1 (red) **h**. A Cox proportional hazards model comparing the survival of the samples with BFB disrupting CCNE1 to samples without BFB disrupting CCNE1 adjusting for age, stage at diagnosis, HRD status and sub-cohort **i** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1186or was performed on all p values. No kegg pathways were found to be significantly ($p<0.05$) enriched in genes with increased or decreased gene expression.

Gene expression and survival impacts of tyfonas

As for rigma and pyrgo, when initially described tyfonas was reported to have significantly worse survival than samples without tyfonas (Hadi et al. 2020), but further investigation is needed. Once the effect of WGD had been accounted for, there were 55 genes with increased expression and 89 genes with decreased expression (Figure 56b). The gene showing one of the largest and most significant expression changes (Figure 56d and 56e) was HIST2H2AA3 which is yet another gene regulated by the PAX8 transcription factor (Adler et al. 2017). The genes that showed most decreased expression in samples with tyfonas were significantly enriched for the KEGG Olfactory transduction pathway (Figure 56f). However, in the combined cohort only seven samples of HGSOC were identified as having tyfonas and the effect on survival did not show a strong effect once age, stage, HRD and cohort were adjusted for (Figure 56g and 56h).


a
60
40
20
0
-log(P value)
−20
−10
0
10
Log2FoldChange
b
40
20
0
-log(P value)
−20
−10
0
10
Log2FoldChange

C

188

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
CACNG2	13.42	9.30	2.11	4.41	3.09E-03
LBX1	10.49	9.07	2.28	3.97	1.44E-02
SLC6A5	32.04	8.95	2.26	3.97	1.44E-02
C20orf173	3.55	7.71	1.75	4.40	3.27E-03
GAGE12H	12.02	7.33	6.40	1.14	9.49E-01
ELF5	55.95	7.22	1.51	4.77	6.48E-04
NKX2-1	7.23	6.44	2.47	2.61	3.63E-01
NEFL	45.57	6.35	1.35	4.71	8.58E-04
AC136616.2	20.89	6.33	2.15	2.94	2.15E-01
MUC2	496.23	6.31	1.80	3.50	6.03E-02

d

Largest decrease in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
AL358075.4	119.89	-24.77	3.43	-7.22	6.53E-10
HIST2H2AA3	1612.41	-23.55	2.87	-8.22	4.56E-13
LRRC55	239.84	-22.31	2.45	-9.11	5.19E-16
AC004556.1	289.19	-21.96	2.83	-7.77	1.25E-11
CTAG2	26.05	-21.68	3.11	-6.97	3.88E-09
AC092143.1	36.92	-21.66	2.41	-8.97	1.40E-15
EPS8L3	44.76	-21.39	2.60	-8.22	4.56E-13
AC012184.2	77.19	-21.32	2.59	-8.23	4.56E-13
MAGEA9	23.95	-21.08	4.74	-4.44	2.73E-03
SSX4B	85.92	-21.00	6.20	-3.39	7.83E-02

e

Most Significant

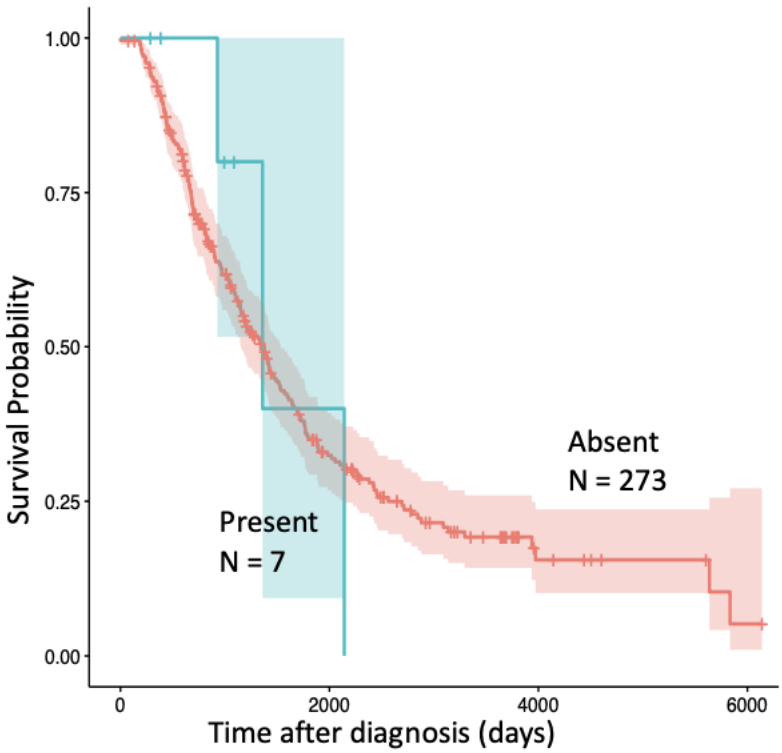
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
RPL36A-HNRNPH2	26.33	-20.91	1.89	-11.08	3.08E-24
BMP5	21.37	-21.00	2.25	-9.33	1.02E-16
LRRC55	239.84	-22.31	2.45	-9.11	5.19E-16
AC092143.1	36.92	-21.66	2.41	-8.97	1.40E-15
DAO	44.45	-16.11	1.80	-8.93	1.65E-15
FDCSP	56.36	-20.35	2.39	-8.51	5.59E-14
HIST2H2AA3	1612.41	-23.55	2.87	-8.22	4.56E-13
EPS8L3	44.76	-21.39	2.60	-8.22	4.56E-13
AC012184.2	77.19	-21.32	2.59	-8.23	4.56E-13
GPR50	18.56	-20.28	2.49	-8.15	6.96E-13

f

Gene Pathway enrichment (Decreased expression)

Pathway	p.adjust	Count	GroupSize
Olfactory transduction	1.34E-07	16	439
Starch and sucrose metabolism	5.02E-02	3	36
Carbohydrate digestion and absorption	7.26E-02	3	47
Alcoholism	9.64E-02	5	187

g



h

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.067	
Stage	N=280	1.55 (1.18-2.03)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.93 (0.69-1.25)	0.638	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.77 (1.20-2.63)	0.004	
	BCCA N=59	0.90 (0.59-1.39)	0.641	
	TCGA N=31	1.66 (1.04-2.63)	0.032	
Tyfonas	Absent N=119	Reference		
	Present N=161	0.86 (0.27-2.78)	0.803	

Figure 56 Gene expression and survival tyfonas

The changes in gene expression between the samples with and without tyfonas corrected for sub-cohort. There were 85 genes with increased expression and 129 genes with decreased expression (**a**)**.** The changes in gene expression between the samples with and without tyfonas corrected for sub-cohort and WGD (**b**) .There were 55 genes with increased expression and 89 genes with decreased expression. Expression data was used from 3 samples with tyfonas and 201 samples without tyfonas. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with tyfonas (blue dots). Negative values represent a decrease in gene expression in samples with tyfonas (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size

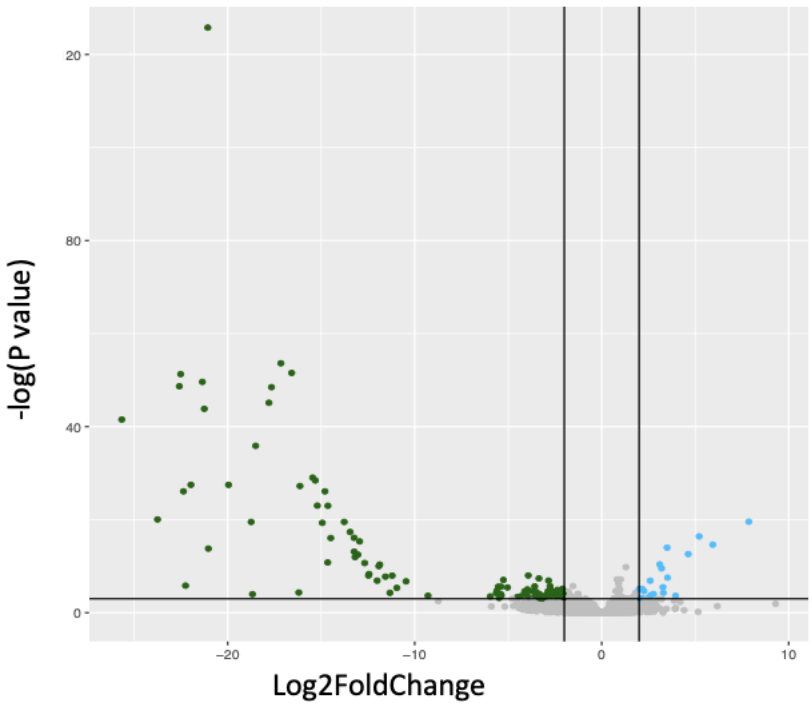
thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with thIreatest increase in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**), the ten genes with the mI significant change in gene expression (**e)**.

Gene set enrichment of genes with decreased expression in samples with the tyfonas using the KEGG terms **f** (M. Kanehisa and Goto 2000; Minoru Kanehisa 2019; Minoru Kanehisa et al. 2023). Kaplan-Meier survival curve for time after diagnosis in days for samples with tyfonas (blue) and without (red) **g**. A Cox proportional hazards model comparing the survival of the samples with tyfonas to samples without tyfonas adjusting for age, stage at diagnosis, HRD status and sub-cohort **h** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochb192or 1995) was performed on all p values. No kegg pathways were found to be significantly ($p<0.05$) enriched in genes with increased gene expression.

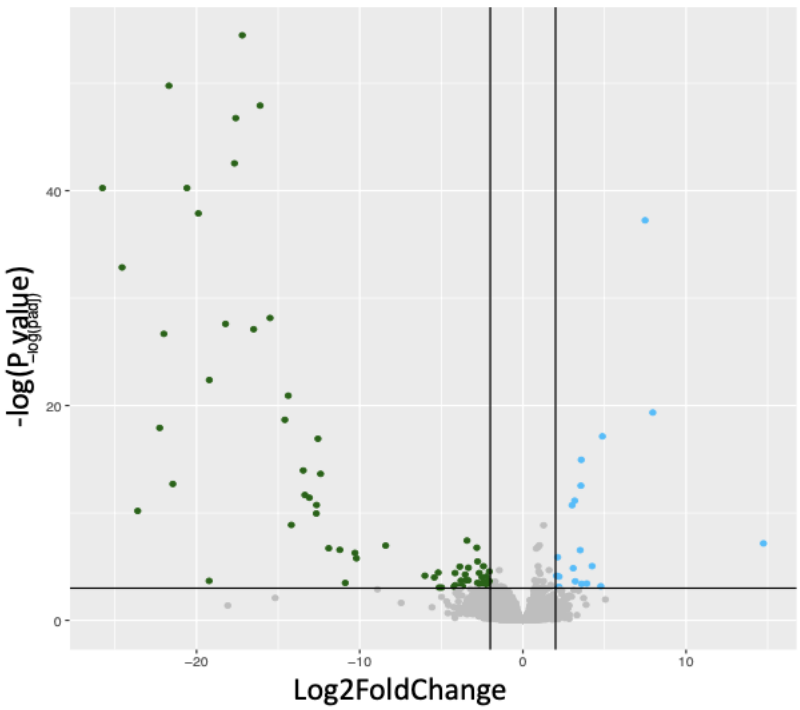
Gene Expression and Survival Impact of Seismic Amplification

When first reported seismic amplification (SA) was stated to increase the gene expression of oncogenes but the effect of SA on patient survival remains unstudied (Rosswog et al. 2021). To investigate the effect of SA on gene expression in HGSOC and its effect on survival samples with and without SA were compared. Overall only 46 genes had significantly increased gene expression and 85 showed a decrease in gene expression in samples with SA (Figure 57b). The only KEGG term enriched in the genes with decreased gene expression was olfactory transduction. There was also no significant effect of SA on survival (Figure 57f and 57g) .

a



b



c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
TBC1D3E	21.36	14.74	3.07	4.81	7.73E-04
OR7G3	17.77	7.96	1.15	6.95	3.95E-09
LCN1	109.07	7.49	0.81	9.21	6.71E-17

SLC17A2	3.91	5.06	1.64	3.09	1.44E-01
BEST3	33.43	4.87	0.74	6.61	3.63E-08
PRAMEF11	36.81	4.77	1.30	3.68	4.16E-02
OR7G2	65.79	4.24	0.99	4.30	6.42E-03
CRNN	27.18	3.92	1.04	3.78	3.25E-02
AC139530.2	14.51	3.87	1.39	2.79	2.34E-01
OR4N2	15.02	3.72	1.17	3.17	1.26E-01

d

Largest decrease in expression

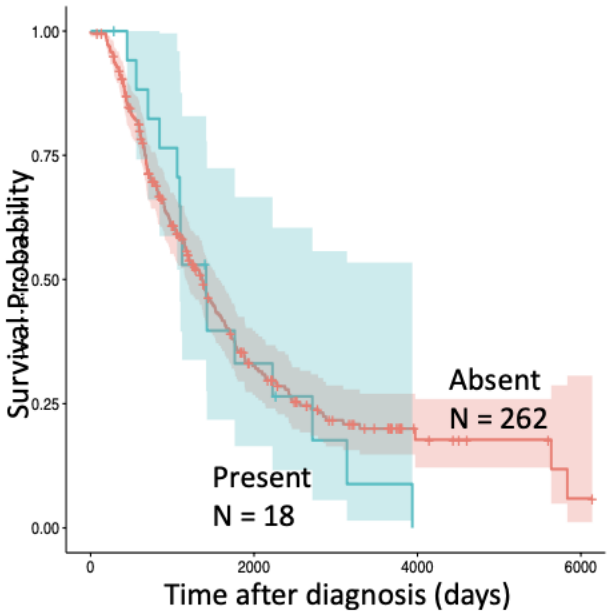
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
AC120057.2	197.78	-25.76	2.70	-9.55	3.32E-18
GAGE12J	72.77	-24.56	2.82	-8.72	5.38E-15
GAGE12C	41.97	-23.60	4.37	-5.40	3.77E-05
SSX4B	85.92	-22.26	3.31	-6.73	1.65E-08
TBC1D3K	14.71	-22.00	2.77	-7.94	2.61E-12
AC011455.2	33.73	-21.69	2.04	-10.61	2.43E-22
GAGE12H	12.02	-21.45	3.65	-5.88	3.06E-06
TRIM49C	14.96	-20.59	2.15	-9.56	3.32E-18
OR6C74	10.20	-19.89	2.14	-9.29	3.52E-17
H2BFS	5.13	-19.23	4.96	-3.88	2.53E-02

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
OR2T6	14.24	-17.20	1.55	-11.10	2.21E-24
AC011455.2	33.73	-21.69	2.04	-10.61	2.43E-22
OR10K2	32.34	-16.11	1.55	-10.40	1.54E-21
OR4K17	9.56	-17.60	1.72	-10.26	4.94E-21
OR9K2	7.35	-17.67	1.80	-9.82	3.33E-19
AC120057.2	197.78	-25.76	2.70	-9.55	3.32E-18
TRIM49C	14.96	-20.59	2.15	-9.56	3.32E-18
OR6C74	10.20	-19.89	2.14	-9.29	3.52E-17
LCN1	109.07	7.49	0.81	9.21	6.71E-17
GAGE12J	72.77	-24.56	2.82	-8.72	5.38E-15

f



g

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.071	
Stage	N=280	1.53 (1.17-2.01)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.94 (0.70-1.26)	0.664	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.84 (1.25-2.73)	0.002	
	BCCA N=59	0.92 (0.60-1.42)	0.717	
	TCGA N=31	1.72 (1.08-2.73)	0.021	
Seismic Amplification	Absent N=262	Reference		
	Present N=18	1.30 (0.76-2.24)	0.337	

Figure 57 Gene expression and survival with seismic amplification

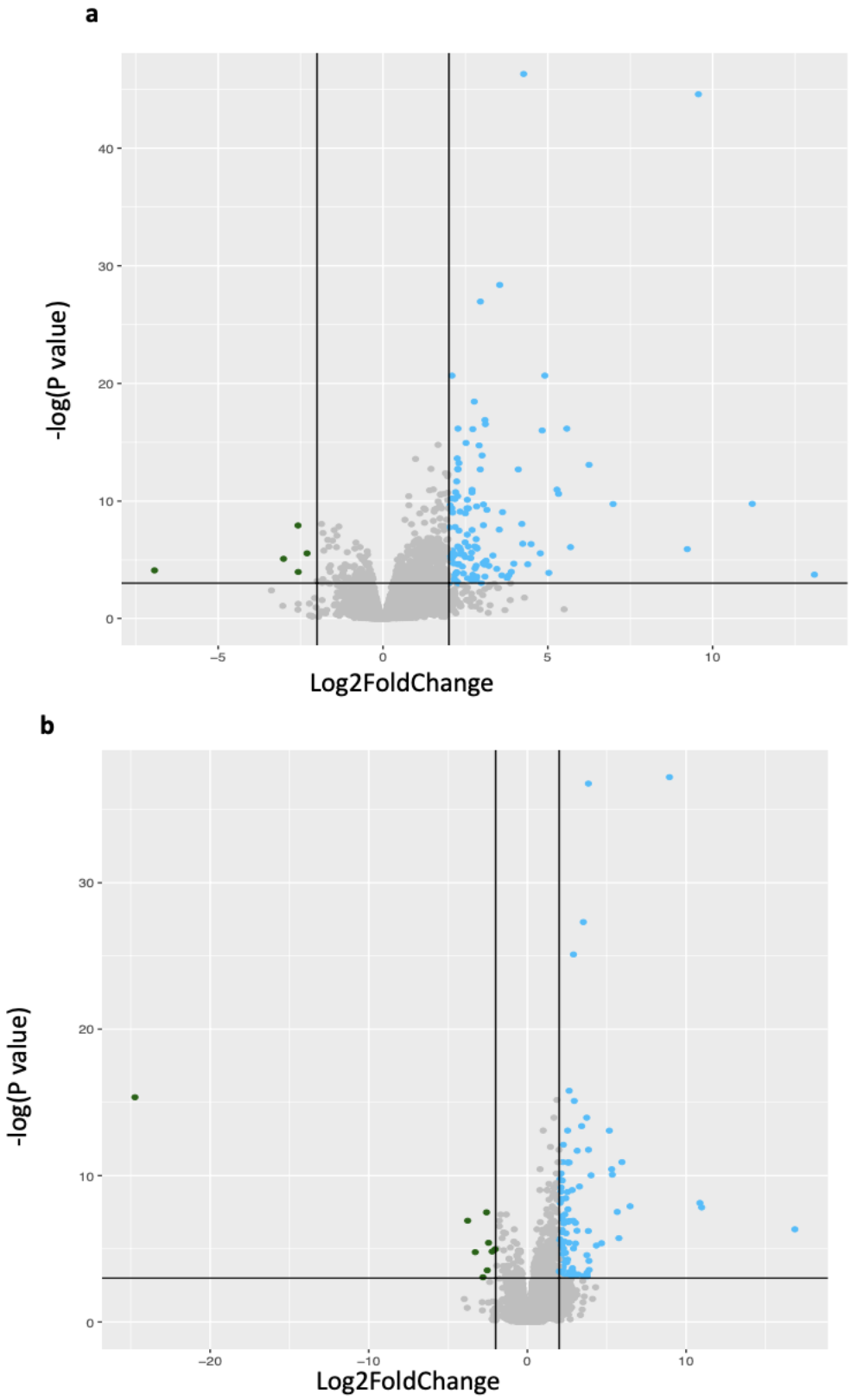
The changes in gene expression between the samples with and without seismic amplification corrected for sub-cohort. There were 22 genes with increased expression and 114 genes with decreased expression (**a**)**.** The changes in gene expression between the samples with and without seismic amplification corrected for sub-cohort and WGD (**b**) .There were 46 genes with increased expression and 85 genes with decreased expression. Expression data was used from 11 samples with seismic amplification and 193 samples without seismic amplification. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with seismic amplification (blue dots). Negative values represent a decrease in gene expression in samples with seismic amplific–tion (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest increase in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**) Olfactory transduction keg term enriched in genes with decreased expression count 28 group size = 439 p=7.80x10-22, the ten genes with thIost significant change in gene expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with seismic amplification (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with seismic amplification to samples without seismic amplification adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values.

Gene Expression and Survival Impacts of ecDNA

The presence of ecDNA in tumour samples is well documented to be associated with the over-expression of oncogenes and worse overall survival in many cancer types (Adelman and Martin 2021; Kim et al. 2020; Robert and Crasta 2022; Shah et al. 2021; Bailey et al. 2020). Once WGD had been adjusted for, 126 genes showed significant increases in

expression and 12 genes had significant decreases in expression in samples with ecDNA. The Kaplan-Meier survival curve suggests that the presence of ecDNA is worse for survival in HGSOC (Figure 58f). However once age, stage, HRD, WGD, and cohort are accounted for there was no significant difference in survival for samples with and without ecDNA. This is consistent with results in Chapter 3, where the majority (70%) of ecDNA was found in the AOCS sub-cohort, which is enriched for patients who did not respond to treatment and had poorer survival (Figure 48). Any effect of ecDNA is therefore confounded by the effect of AOCS cohort membership.

To try test the effect of ecDNA on survival without the confounding cohort effect, the impact of ecDNA on survival was investigated in samples within the AOCS sub-cohort only, comparing those with and without ecDNA. Although survival appeared to be worse in AOCS samples with ecDNA (Figure 58h) it did not reach statistical significance when age, stage, HRD and WGD were adjusted for (Figure 58i). Only 14 ecDNA samples were identified in the combined cohort, excluding AOCS. The survival rate of these non-AOCS samples also failed to show a significant difference when compared to the rest of the combined cohort excluding AOCS (Figure 58j and 58k). There is therefore no convincing evidence for an association between ecDNA and survival in the current data, beyond the strong cohort bias (enriched ecDNA in AOCS) I have described.


a
-log(P value)
40
30
20
10
0
−5
0
5
10
Log2FoldChange
b
-log(P value)
30
20
10
0
−20
−10
0
10
Log2FoldChange

c

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
PRAMEF26	0.14	16.86	3.79	4.45	1.79E-03
SSU72P4	2.57	10.98	2.26	4.85	4.00E-04
MBD3L5	1.77	10.87	2.21	4.92	2.94E-04
AL358075.4	119.89	8.96	0.95	9.45	6.89E-17
MBD3L2	8.88	6.47	1.33	4.87	3.70E-04
MYF5	6.68	5.96	1.06	5.62	1.81E-05
SPATA31D4	64.01	5.77	1.35	4.28	3.27E-03
TRIM49D1	35.88	5.66	1.19	4.78	5.44E-04
POTEB2	16.39	5.35	0.99	5.41	4.27E-05
XAGE1A	110.60	5.31	0.97	5.49	2.94E-05

d

Largest decrease in expression

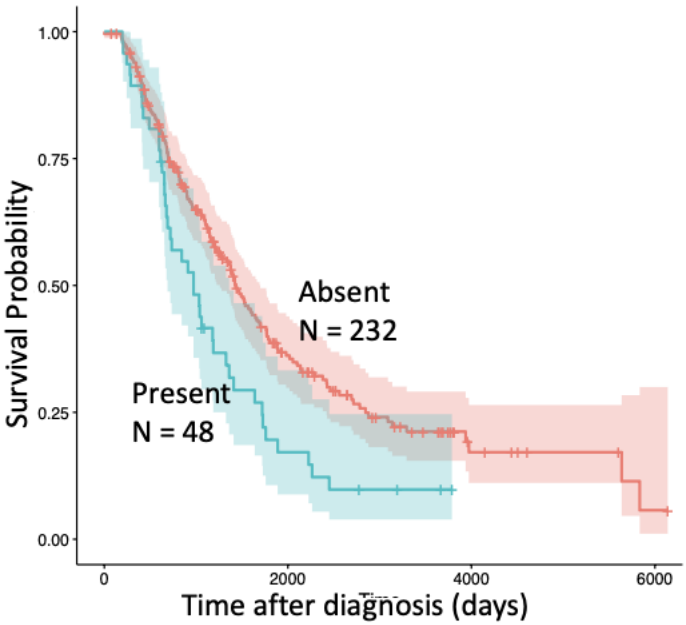
GeneName	baseMean	log2FoldChange	lfcSE	stat	p
TSPY3	5.04	-24.74	3.79	-6.53	2.16E-07
OR51A2	9.75	-3.98	1.63	-2.44	2.09E-01
GDF1	39.68	-3.79	1.96	-1.94	3.81E-01
SSX5	19.96	-3.76	0.81	-4.62	9.88E-04
MAGEA6	69.11	-3.29	0.84	-3.93	8.48E-03
OR5M1	7.10	-2.85	1.26	-2.27	2.58E-01
USP17L3	1.52	-2.84	1.62	-1.75	4.52E-01
MAGEA3	50.85	-2.80	0.86	-3.27	4.71E-02
CRISP3	332.60	-2.58	0.54	-4.76	5.60E-04
MAGEA12	17.41	-2.53	0.72	-3.49	2.94E-02

e

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
AL358075.4	119.89	8.96	0.95	9.45	6.89E-17
PNMA5	43.75	3.84	0.41	9.33	1.07E-16
SLC7A14	31.72	3.53	0.43	8.22	1.37E-12
NUP210L	47.06	2.90	0.37	7.91	1.25E-11
KCNF1	137.24	2.63	0.40	6.62	1.38E-07
TSPY3	5.04	-24.74	3.79	-6.53	2.16E-07
ALS2CR12	135.17	1.86	0.29	6.48	2.59E-07
PPEF2	9.94	2.95	0.46	6.45	2.77E-07
TTBK1	74.97	1.67	0.27	6.24	8.70E-07
BTBD17	33.49	3.74	0.60	6.25	8.70E-07

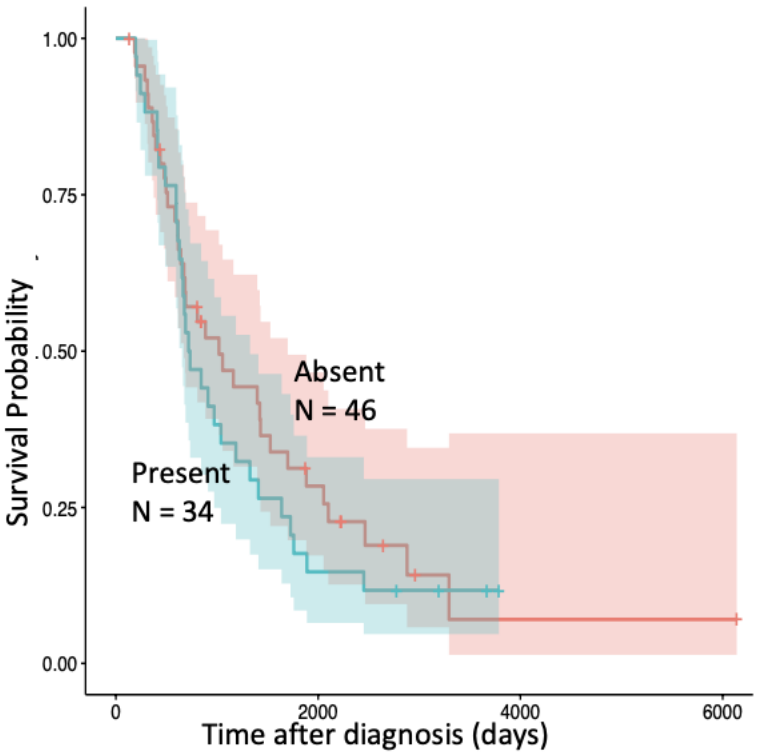
f



g

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.071	
Stage	N=280	1.53 (1.17-2.01)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.94 (0.70-1.26)	0.664	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.84 (1.25-2.73)	0.002	
	BCCA N=59	0.92 (0.60-1.42)	0.717	
	TCGA N=31	1.72 (1.08-2.73)	0.021	
EcDNA	Absent N=262	Reference		
	Present N=18	1.01 (0.67-21.52)	0.954	

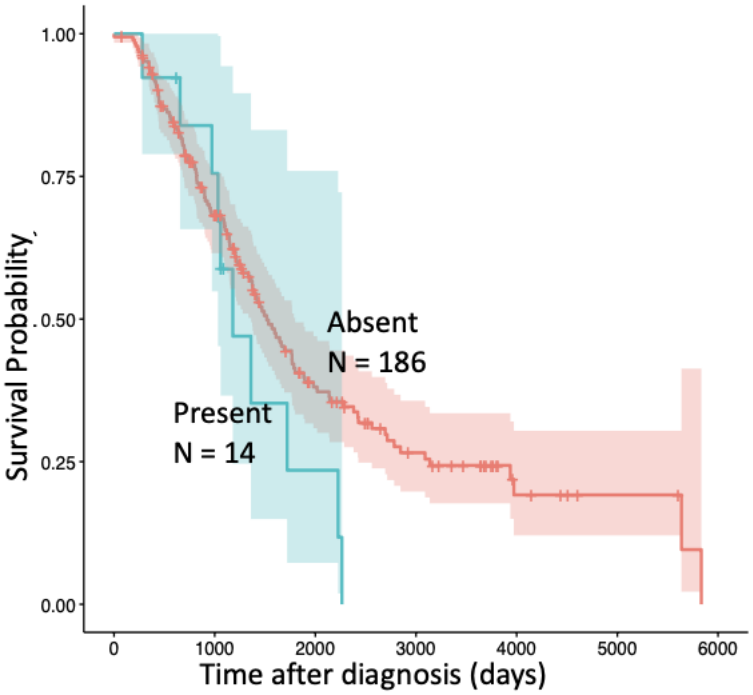
h



i

Condition		Hazard ratio	P value	Forest plot
Age	N=80	1.01 (0.98-1.02)	0.416	
Stage	N=80	0.51 (0.23-1.1)	0.101	
HRD	Absent N=24	Reference		
	Present N=56	0.67 (0.38-1.2)	0.154	
WGD	Absent N=33	Reference		
	Present N=47	1.12 (0.66-1.9)	0.678	
EcDNA In AOCS	Absent N=46	Reference		
	Present N=34	1.11 (0.68-1.8)	0.674	

j



k

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.02 (1.00-1.04)	0.049	
Stage	N=280	2.19 (1.56-3.09)	0.001	
HRD	Absent N=118	Reference		
	Present N=162	0.35 (0.24-0.53)	0.001	
WGD	Absent N=140	Reference		
	Present N=140	0.87 (0.60-1.25)	0. 454	
Cohort	SHGSOC N=110	Reference		
	BCCA N=59	0.86 (0.55-1.35)	0.515	
	TCGA N=31	1.52 (0.96-2.42)	0.077	
EcDNA	Absent N=262	Reference		
	Present N=18	0.79 (0.40-1.59)	0.514	

Figure 58 Gene expression and survival ecDNA

The changes in gene expression between the samples with and without ecDNA corrected for sub-cohort. There were 148 genes with increased expression and 8 genes with decreased expression (**a**)**.** The changes in gene expression between the samples with and without ecDNA corrected for sub-cohort and WGD (**b**) .There were 126 genes with increased expression and 12 genes with decreased expression. Expression data was used from 46 samples with ecDNA and 158 samples without seismic amplification. For both **a** and **b** the X axis shows the log2 fold change with positive values representing an increase of gene expression in samples with ecDNA (blue dots). Negative values represent a decrease in gene expression in samples with ecDNA (green dots). The Y axis represent the -log10 p-values. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are gray. Changes in gene expression was assessed using the DESeq2 package(Love, Huber, and Anders 2014)**.** Summary statistics for the change in gene expression shown in **b**; the ten genes with the greatest increase in gene expression (**c**), the ten genes with the greatest decrease in gene expression (**d**) Olfactory transduction keg term enriched in genes with decreased expression count 28 group size = 439 p=7.80x10-22, the ten genes with the most significant change in gene expression (**e)**. Kaplan-Meier survival curve for time after diagnosis in days for samples with ecDNA (blue) and without (red) **f**. A Cox proportional hazards model comparing the survival of the samples with ecDNA to samples without ecDNA adjusting for age, stage at diagnosis, HRD status and sub-cohort **g** a hazard ratio of 1 (no effect) is shown by a dashed line. Kaplan-Meier survival curve for time after diagnosis in days for samples with ecDNA in the AOCS sub-cohort (blue) and without ecDNA in the AOCS sub-cohort (red) **h**. A Cox proportional hazards model comparing the survival of the samples with ecDNA in the AOCS sub-cohort to samples without ecDNA in the AOCS sub-cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort **I** a hazard ratio of 1 (no effect) is shown by a dashed line. Kaplan-Meier survival curve for time after diagnosis in days for samples with ecDNA in the combined cohort excluding AOCS sub-cohort (blue) and without ecDNA in combined cohort excluding AOCS sub-cohort (red) **j**. A Cox proportional

hazards model comparing the survival of the samples with ecDNA in the in the combined cohort excluding AOCS sub-cohort to samples without ecDNA in the combined cohort excluding AOCS sub-cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort **k** a hazard ratio of 1 (no effect) is shown by a dashed line. Multiple testing correction (Benjamini and Hochberg, 1995) was performed on all p values.

The impact on Gene Expression and Overall Survival of Genomic Instability Measured by Number of SVs

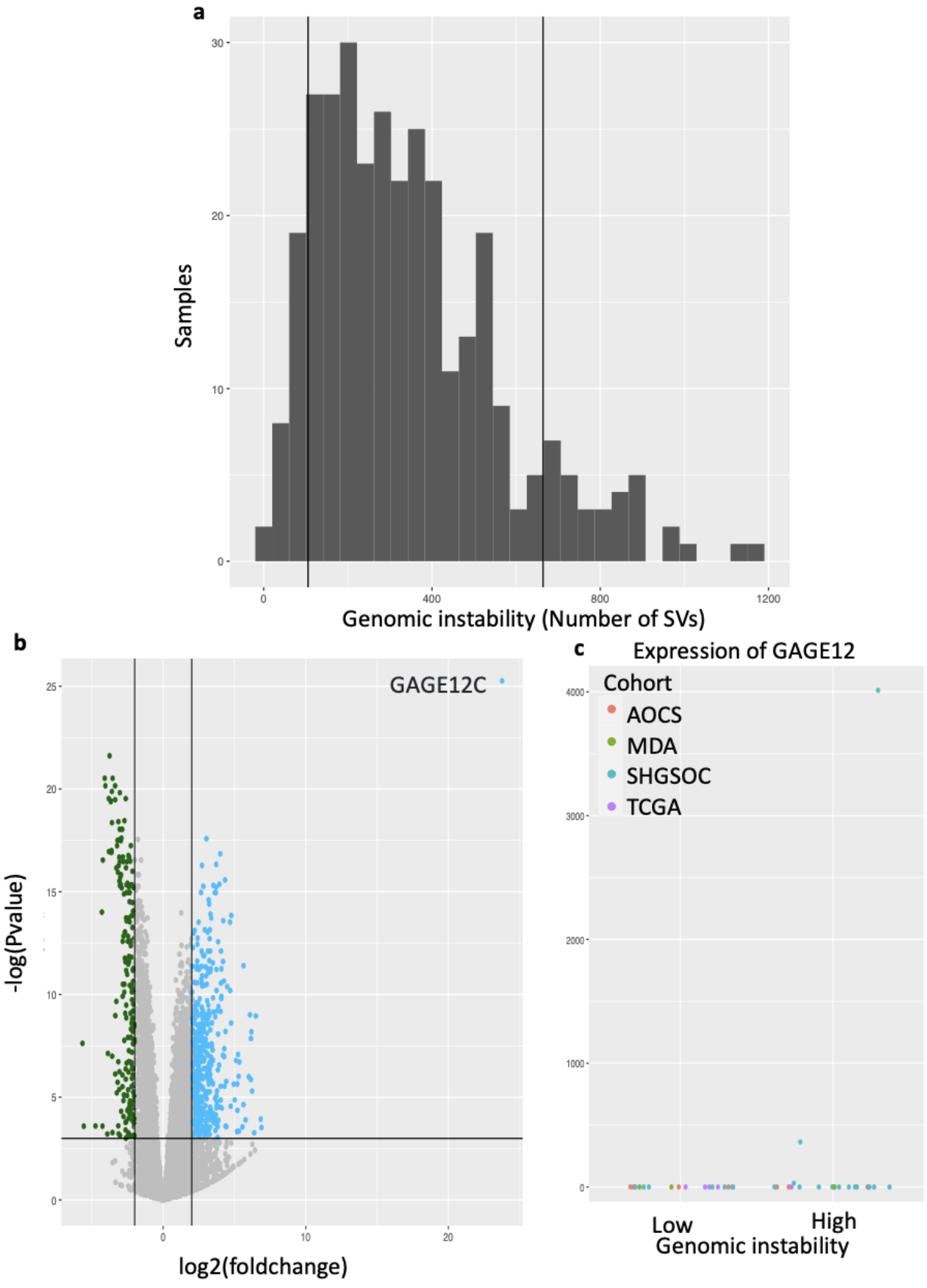
In a study of 126 HGSOC samples, it was reported that samples with a higher number of SVs had better survival time after diagnosis (Garsed et al. 2022). To determine if this trend is also present in a larger cohort of 324 HGSOC samples, the range of genomic instability was measured by the number of SVs (Figure 59a).

To investigate how genomic instability, as measured by the number of SVs, affects gene expression, the most genomically unstable 10% of samples were compared to the most genomically stable 10%. In this comparison 662 genes showed increased expression and 205 genes had decreased expression in the most genomically unstable samples (Figure 59b). Notably, the gene with the largest increase in expression was GAGE12C, a member of the GAGE protein family of testis antigens that are frequently upregulated in cancers and have been proposed as possible therapeutic targets (Gjerstorff and Ditzel 2008). However, despite being significantly increased in expression, this increase was driven by only three samples (Figure 59c).

The impact of genomic instability, as measured by the number of SVs, on survival was assessed by comparing the survival time after diagnosis of the 10% most genomically unstable samples with the 10% most genomically stable samples, in comparison to the rest of the combined cohort. The trend suggests that the most genomically unstable samples had better survival than the rest of the combined cohort, and the most genomically stable samples had worse survival than the rest of the combined cohort (Figure 59i). However, neither of these trends reach statistical significance (Figure 59j).

To explore the sensitivity of this trend to the threshold used to define genomic instability, the threshold for identifying the most genomically unstable samples was varied. The thresholds used were: the top percentile of samples above 50%, 75%, 90%, and 95% of the other samples. The trend of the most genomically unstable samples having better survival (Hazard ratios < 1) is constant across all thresholds tested but

does not reach statistical significance (Figure 59k and 59l).



d

Largest increase in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GAGE12C	10.38	23.77	2.94	8.09	1.07E-11
OR14A2	10.73	6.89	2.52	2.73	2.94E-02
Z83844.1	6.84	6.84	2.35	2.91	1.96E-02
TPTE	23.35	6.50	1.40	4.63	1.30E-04
DEFB108B	5.17	6.43	2.94	2.19	8.84E-02
FOXR2	5.62	6.38	2.44	2.62	3.82E-02
OR14A16	7.04	6.25	2.67	2.34	6.66E-02
OR1C1	11.24	6.23	1.81	3.45	5.01E-03
TRIM6-TRIM34	17.28	6.19	1.40	4.41	2.80E-04
CTAG1A	38.10	6.16	1.43	4.31	3.86E-04

e

Largest decrease in expression

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
AC004556.1	161.21	-5.66	1.34	-4.24	4.90E-04
AC034102.2	352.04	-5.57	2.02	-2.76	2.76E-02
GAGE2E	6.96	-4.75	1.72	-2.77	2.73E-02
RPSA	242643.20	-4.30	0.73	-5.88	8.25E-07
KRTAP5-2	4.40	-4.26	1.54	-2.76	2.75E-02
CXCL17	1095.10	-4.23	0.66	-6.43	6.57E-08
RPS2	244466.70	-4.10	0.56	-7.31	1.22E-09
RPL31	112495.70	-4.06	0.56	-7.21	1.77E-09
AC011511.4	222.35	-3.92	1.51	-2.59	4.04E-02
SLC13A2	15.89	-3.88	0.95	-4.08	7.99E-04

f

Most Significant

GeneName	baseMean	log2FoldChange	lfcSE	stat	p
GAGE12C	10.38	23.77	2.94	8.09	1.07E-11
RPL5	105200.44	-3.76	0.50	-7.55	4.11E-10
RPS2	244466.73	-4.10	0.56	-7.31	1.22E-09
RPS26	12290.91	-3.55	0.48	-7.32	1.22E-09
RPL31	112495.72	-4.06	0.56	-7.21	1.77E-09
RPL32	66849.75	-3.37	0.47	-7.23	1.77E-09
RPL13A	203938.38	-3.60	0.50	-7.15	2.33E-09
RPL6	79682.52	-3.04	0.43	-7.12	2.47E-09
MT-ND6	29816.72	-3.82	0.54	-7.05	3.29E-09
NACA	32228.05	-2.63	0.37	-7.06	3.29E-09

g

Gene Pathway enrichment (Increased expression)

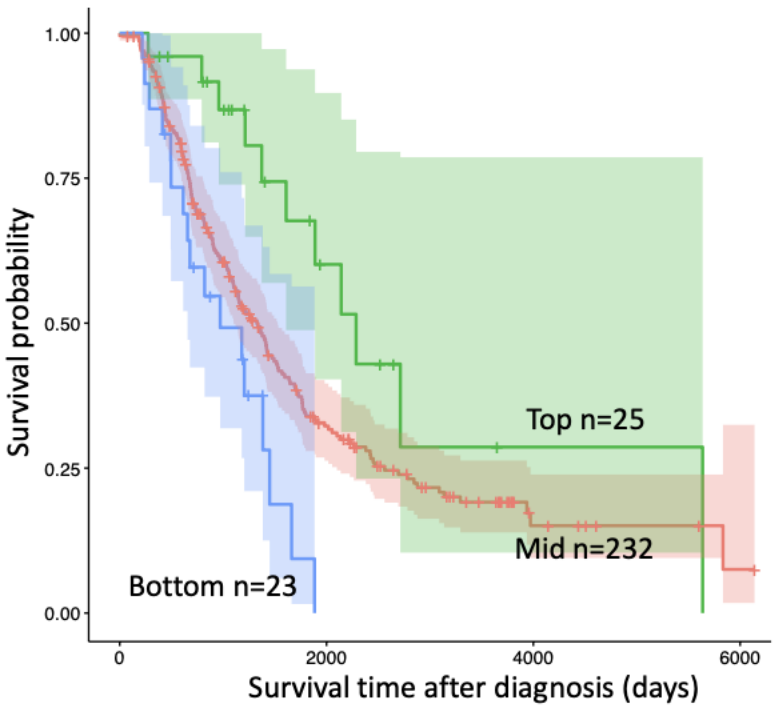
Pathway	p.adjust	Count	GroupSize
Olfactory transduction	2.92E-13	52	439
Neuroactive ligand-receptor interaction	2.05E-11	44	367
cAMP signaling pathway	2.07E-04	23	225
Amphetamine addiction	2.07E-04	12	69
Nicotine addiction	3.19E-04	9	40
Calcium signaling pathway	3.19E-04	23	240
Taste transduction	1.22E-03	12	86

h

Gene Pathway enrichment (Decreased expression)

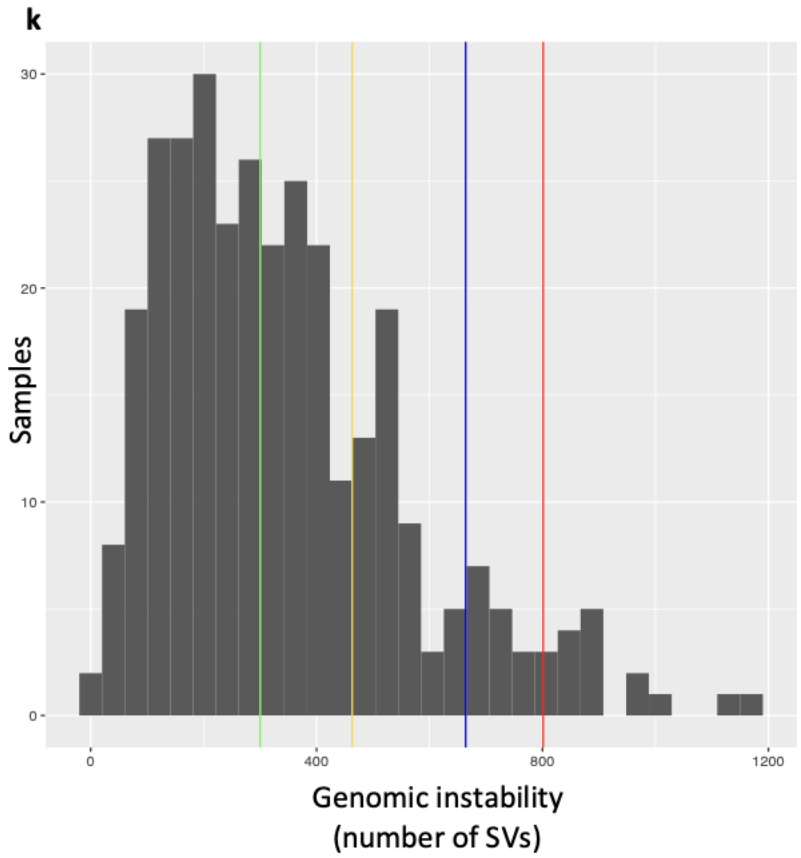
Pathway	p.adjust	Count	GroupS–ze
Ribosome	9.45E-75	62	167
Coronavirus disease -—COVID-19	1.23E-64	62	232
Chemical carcinogenesis - reactive oxygen species	1.49E-02	11	223
Parkinson disease	1.49E-02	12	266
Oxidative phosphorylation	1.85E-02	8	134
Thermogenesis	4.24E-02	10	232

i



j

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.02 (1.00-1.03)	0.071	
Stage	N=280	1.49 (1.13-2.00)	0.002	
HRD	Absent N=118	Reference		
	Present N=162	0.51 (0.36-0.70)	0.001	
Cohort	SHGSOC N=110	Reference		
	AOCS N=80	1.84 (1.25-2.73)	0.008	
	BCCA N=59	0.92 (0.60-1.42)	0.471	
	TCGA N=31	1.72 (1.08-2.73)	0.059	
Genomic instability	middle N=232	Reference		
	High N=25	0.65 (0.34-1.20)	0.188	
	Low N=23	1.38 (0.80-2.4)	0.242	


k
30
20
10
0
Samples
0
400
800
1200
Genomic instability
(number of SVs)

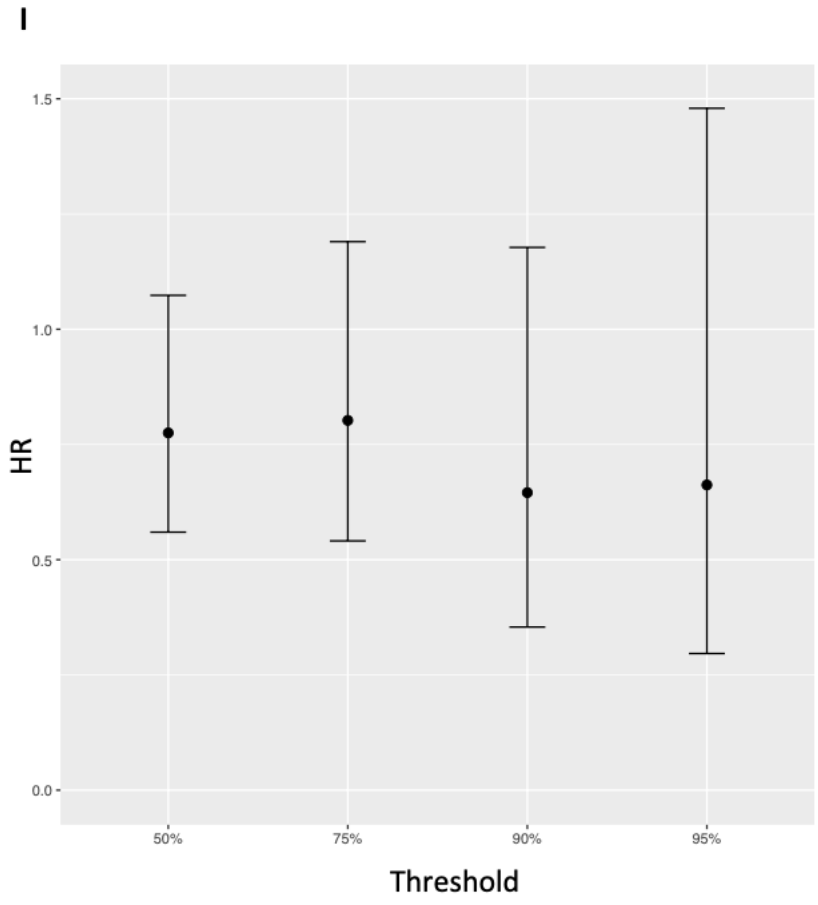

l
1.5
1.0
0.5
0.0
HR
50%
75%
90%
95%
Threshold

Figure 59 Genomic instability (SV burden) impact on gene expression and overall survival

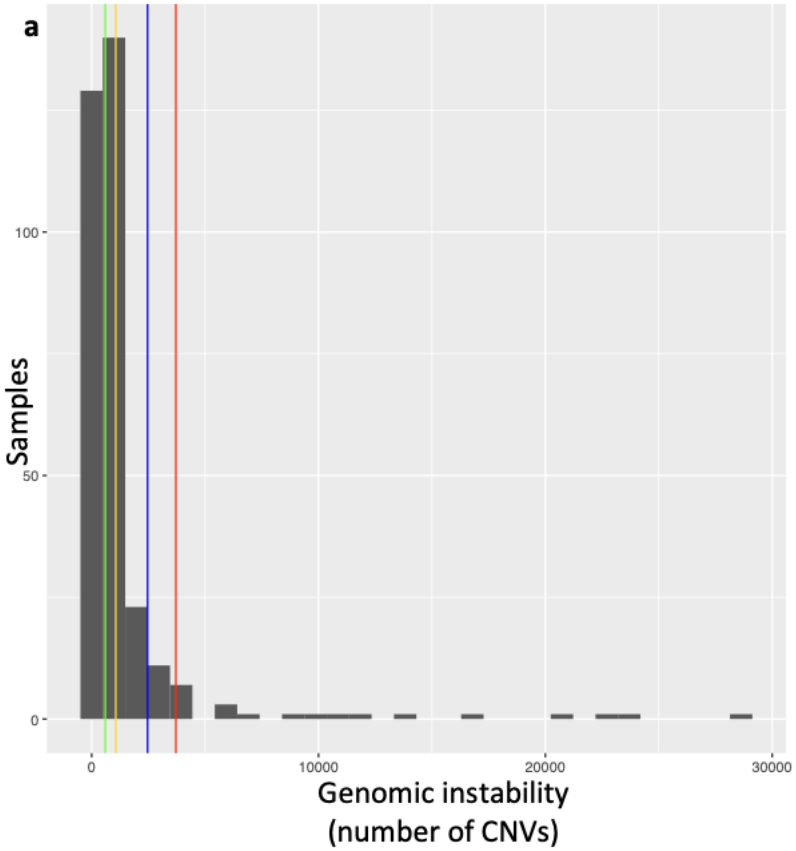
The distribution of genomic instability with vertical lines showing the threshold for the top and bottom 10% of genomic instability a. The change in gene expression between the top 10% most genomically unstable samples and the 10% least genomically unstable samples based on quantity of SV in the samples b. The X axis shows the log2 fold change with positive values representing an increase of gene expression in more genomically unstable samples (blue dots). Negative values represent a decrease in gene expression in more genomically unstable samples (green dots). There were 662 genes upregulated and 205 genes downregulated. 21 samples in the top 10% of genomically unstable samples had gene expression data. 17 samples in the bottom 10% of genomically unstable samples had gene expression data. The Y axis represent the -log10 p-values after multiple testing correction (Benjamini and Hochberg, 1995) method. The horizontal line represents the threshold for significance with values above that line being significant. The two vertical lines represent a fold change in expression of -2 and 2. Genes that did not pass the significance and effect size thresholds are grey. Changes in gene expression was assessed using the 212ormalizedkage (Love, Huber, and Anders 2014). The normalised read counts of the GAGE12C for the most and least genomically instable samples c.

Summary statistics for the change in gene expression shown in b; the ten genes with the greatest increase in gene expression (d), the ten gIs with the greatest decrease in gene expression (e), the ten genes with the most significant change in gene expression (f). Gene set enrichment of genes with increased expression in samples with the highest genomic instability using the KEGG terms g (M. Kanehisa and Goto 2000; Minoru Kanehisa 2019; Minoru Kanehisa et al. 2023). Gene set enrichment of genes with decreased expression in samples with the highest genomic instability using the KEGG terms h (M. Kanehisa and Goto 2000; Minoru Kanehisa 2019; Minoru Kanehisa et al. 2023). A Kaplan-Meier survival curve for time after diagnosis for three groups the highest genomic instability (green), the least genomic instability (blue), and remainder of the combined cohort (red) i. A Cox proportional hazards model comparing the survival of the most and least genomically unstable samples to the remainder of the combined

cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort j. The distribution of genomic instability as measured by number of SV in the sample with four lines show the quartile for 50% (green), 75% (yellow), 90% (blue), 95% (red) k. The hazard ratio for a Cox proportional hazards model comparing the samples above and below the respective percentiles (50%,75%,90%,95%) of genomically instable samples to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort l. A hazard of 1 represent no effect and hazard ratios of less than 1 represent improved survival.

The Impact on Gene Expression and Overall Survival of Genomic Instability Measured by Number of CNVs

To examine whether the trend of increased genomic instability is also associated with better survival when genomic instability as measured by the number of CNVs, four thresholds were used, with genomic instability defined as greater than the percentiles 50%, 75%, 90%, and 95% of the other samples. The trend of the most genomically unstable samples having better survival (Hazard ratios < 1) is barely present across the thresholds tested and does not reach statistical significance (Figure 60a and 60b).



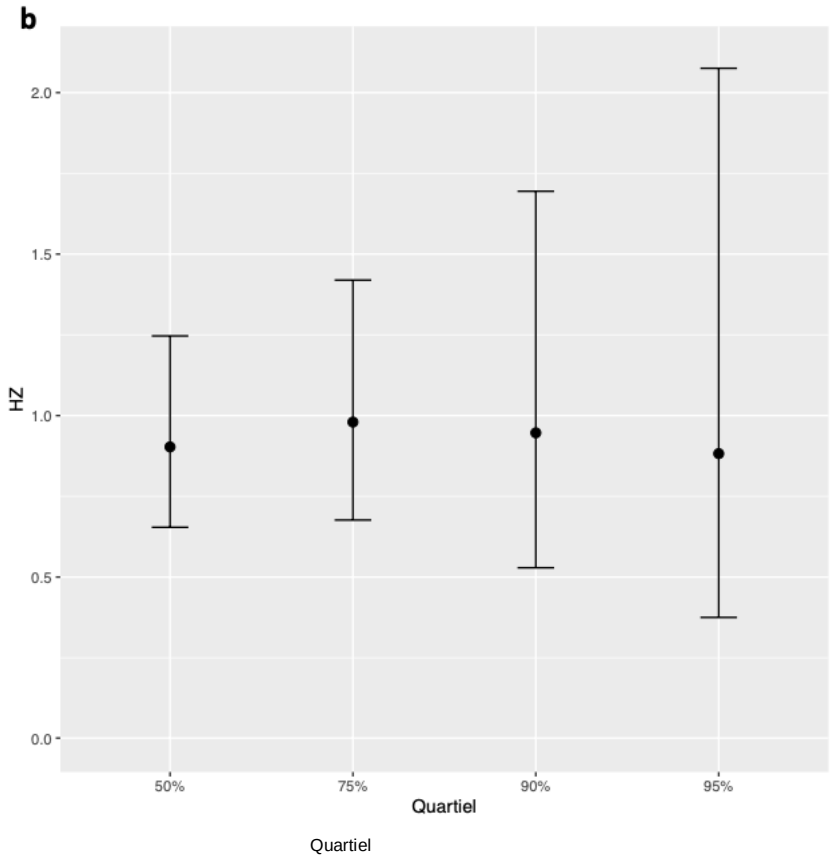


Figure 60 Genomic instability measured by CNV burden

The distribution of genomic instability as measured by the number of CNVs in the sample with four lines showing for 50% (green), 75% (yellow), 90% (blue), 95% (red) **a**. The hazard ratio for a Cox proportional hazards model comparing the samples above and below the respective quartile (50%,75%,90%,95%) of genomically unstable samples to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort **b**. A hazard ratio of 1 indicates no effect, and a hazard ratio less than 1 indicates improved survival.

Severity of Events

In chapter 3, it is shown that cSVs contain a range of number of SVs within a cSV, as well as a ranges of size. To investigate if cSV at the extremes of the spectrum of severity (number of SV explained or size of the cSV) have an impact on survival, the top 25% most severely affected and the bottom 25% severely affected will be compared to the

rest of the samples with that cSV.

The samples with the least severe chromoplexy as measured by number of SVs explained, showed the worst survival (Figure 61 b). However, this does not reach statistical significance when modelled (Figure 61 c). As chromoplexy consists of chains of translocation it does not have a biologically meaningful size.

When measured by the number of structural variations (SVs), the least severe samples of breakage-fusion-bridge (BFB) had a significantly worse survival rate than the rest of the combined cohort, with more than three times the risk of death (Figure 62 b and c). Moreover, the most severe samples of BFB also had an increased risk of death compared to the rest of the combined cohort, although this difference did not reach statistical significance.

When the size of the BFB was used as a measure of severity, the least severe samples also showed significantly worse survival, with an increased risk of death of 2.33 compared to the rest of the combined cohort (Figure 62 e and f). Similarly, the most severe samples had an increased risk of death of 1.72 times compared to the rest of the combined cohort, but again, this difference did not reach statistical significance.

The most and least severe rigma samples did not show a clear impact on survival when severity was measured by number of explained SVs (Figure 63 b and c). A similar trend was seen when severity was measured by size of cSV. However, the most severe rigma samples had increased risk of death (1.79) compared to the rest of the combined cohort but this failed to reach statistical significance (Figure 63 e and f).

The most and least severe pyrgo samples did not show a clear impact on survival when severity was measured by number of explained SVs (Figure 64 b and c). When severity was measured by size of cSV both the most and least severe prygo samples had decreased risk of death (0.85) compared to the rest of the combined cohort but this failed to reach statistical significance as well (Figure 64 e and f).

The chromothripsis samples with the most severe impact, when measured by the number of explained SVs, showed better survival than the rest of the combined cohort, with less than half the risk of death (0.49) compared to the rest of the cohort. However, this difference did not reach statistical significance (Figure 65 b and c). On the other hand, the chromothripsis samples with the least severe impact had significantly worse survival than the rest of the combined cohort, with an increased risk of death of 2.45 compared to the rest of the cohort (Figure 65 b and c).

The most severely impacted chromothripsis samples when measured by size of chromothripsis, showed significantly better survival than the rest of the combined cohort and had a decreased risk of death with only 0.31 times the risk of death of the combined cohort (Figure 65 e and f). However, the least severely impacted chromothripsis samples did not show a significant impact on survival and had a decreased risk of death of 0.78 times compared to the rest of the combined cohort.

When measured by number of explained SVs, the most and least severe ecDNA samples did not show a clear impact on survival compared to the combined cohort (Figure 66 b). However, after adjusting for age, stage, HRD, and cohort, the most severely impacted ecDNA samples showed a decreased risk of death (0.55) compared to the rest of the combined cohort, but this difference did not reach statistical significance (Figure 66 c). The least severely impacted ecDNA samples had increased risk of death (1.40) but this was also not significant (Figure 66 c).

When measured by size of ecDNA the most and least severe samples both had better survival than the rest of the combined cohort (Figure 66 e). The improved survival reaches statistical significance for the most severe ecDNA sample with a decreased risk of death of only 0.13 compared to the combined cohort (Figure 66 f).

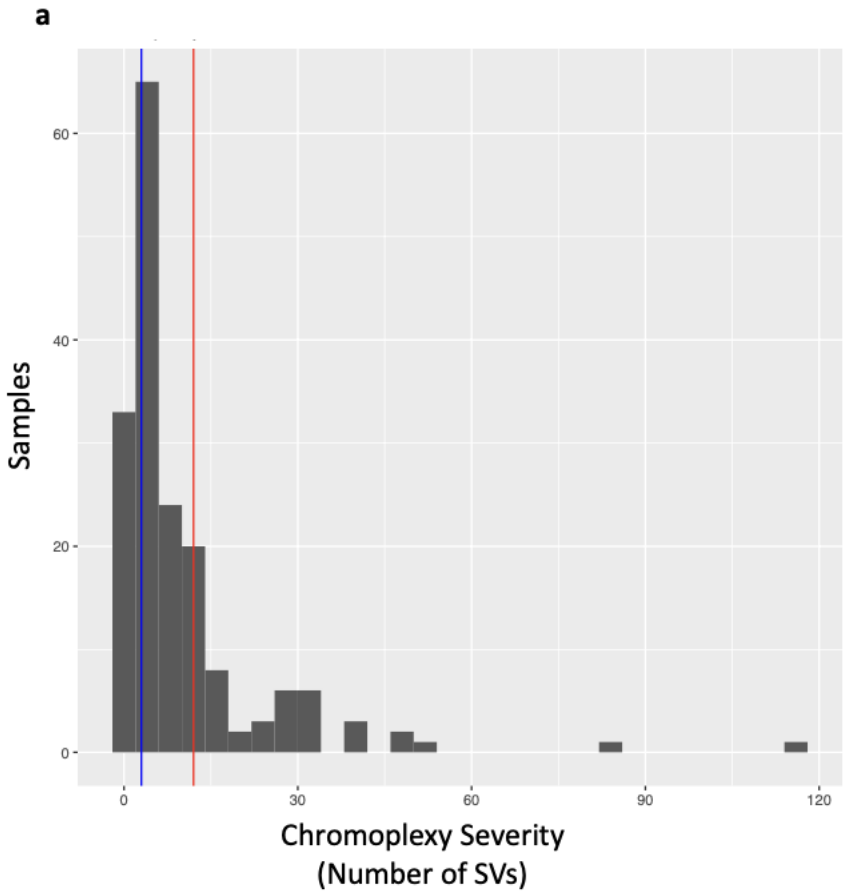
To investigate the survival impact of cSV severity, regardless of cSV type, the most severely impacted samples were compared to the rest of the–samples that contained at least one cSV (Figure 67 b - Figure 67 f). Four thresholds were initially test to define most severely impacted samples. These thresholds were as follows 50%,75%,90% and 95% of samples containing at least one cSV (Figure 67a). The resulting hazard ratio for the most severe cSV for each threshold tested showed that there is a general trend that the most severe cSV samples have lower risk (Hazard ratio < 1) compared to the rest of the combined cohort (Figure 67 b) It is hard to observe if a difference in survival in the Kaplan-Meier survival curve is significant (Figure 67 c). However, once age, stage, HRD and cohort have been corrected for being in the top 25% most severe cSV samples measured by number of SV is significantly better for survival (Figure 67 d).

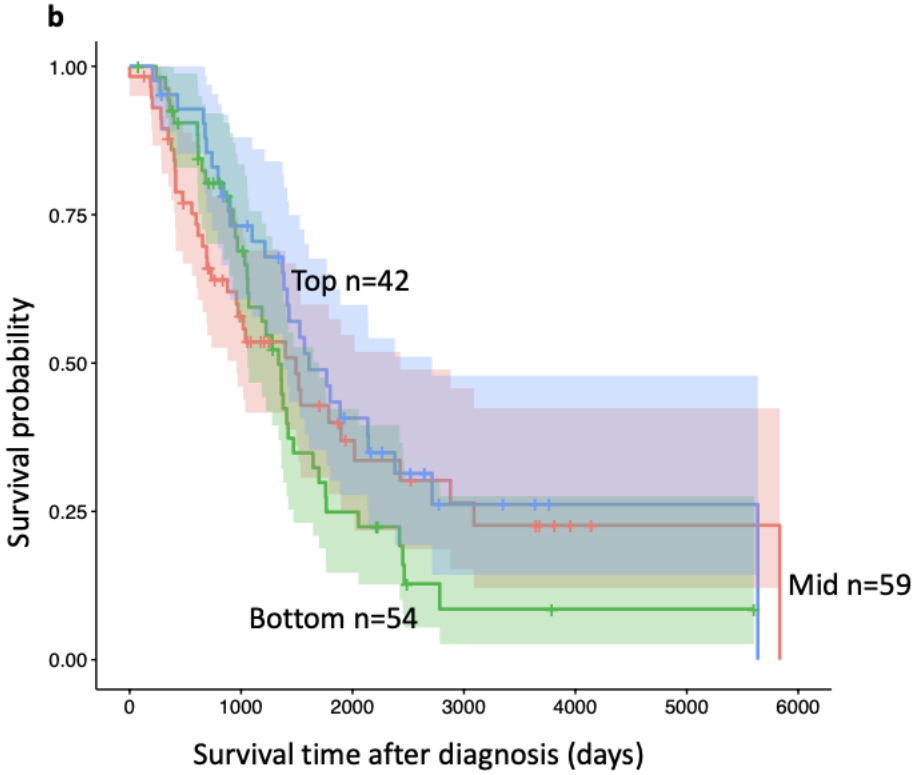
Four thresholds were initially tested to define most severely impacted samples by cSV size. These samples had cSV larger than 50%,75%,90% and 95% of samples containing at least one cSV (Figure 67 e). The trend across all threshold of severity measured by total size of cSV in samples is that the most severe samples have decreased risk of death (hazard ratio < 1) compared to the rest of the combined cohort (Figure 67 f). The Kaplan-Meier survival curve shows the top 25% most severely effect samples by size of cSV in the sample have significantly better survival than the rest of the combined cohort (Figure 67 g and h).

In chapter 4, it was shown that the majority of clusters of SV remain unexplained by cSV. To investigate if the severity of cluster in a sample measured by number of clusters impacts survival, the top 25% most severely impacted and the bottom 25% least severely impacted samples were compared to the rest of the combined cohort. The

most severely impacted samples had the best survival compared to the rest of the combined cohort but this did not reach statistical significance (Figure 68 b and c).

To measure severity, the total size of clusters was used. I compared the survival of the top 25% most severely impacted samples and the bottom 25% least severely impacted samples to the rest of the combined cohort. The least severe samples had the worse survival compared to the rest of the combined cohort but this did not reach statistical significance (Figure 68 e and h f).


a
Samples
60
40
20
0
0
30
60
90
120
Chromoplexy Severity
(Number of SVs)

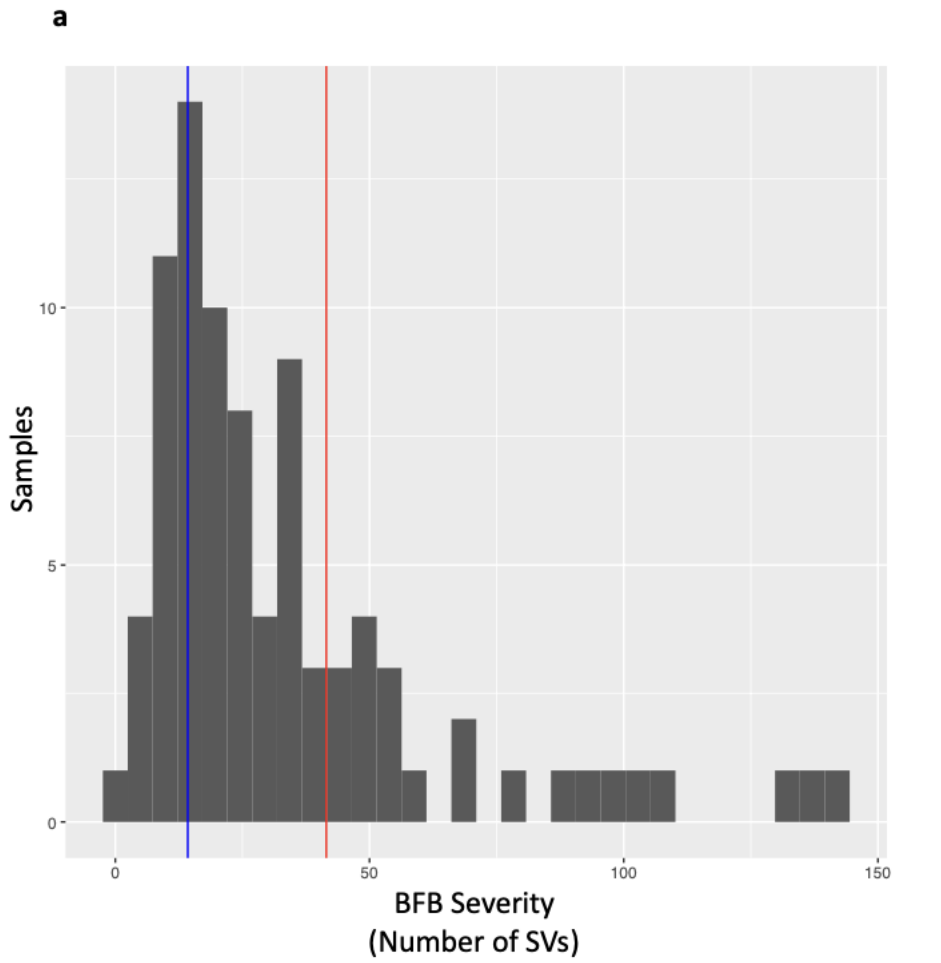

b
Survival probability
1.00
0.75
0.50
0.25
0.00
Top n=42
Mid n=59
Bottom n=54
0
1000
2000
3000
4000
5000
6000
Survival time after diagnosis (days)

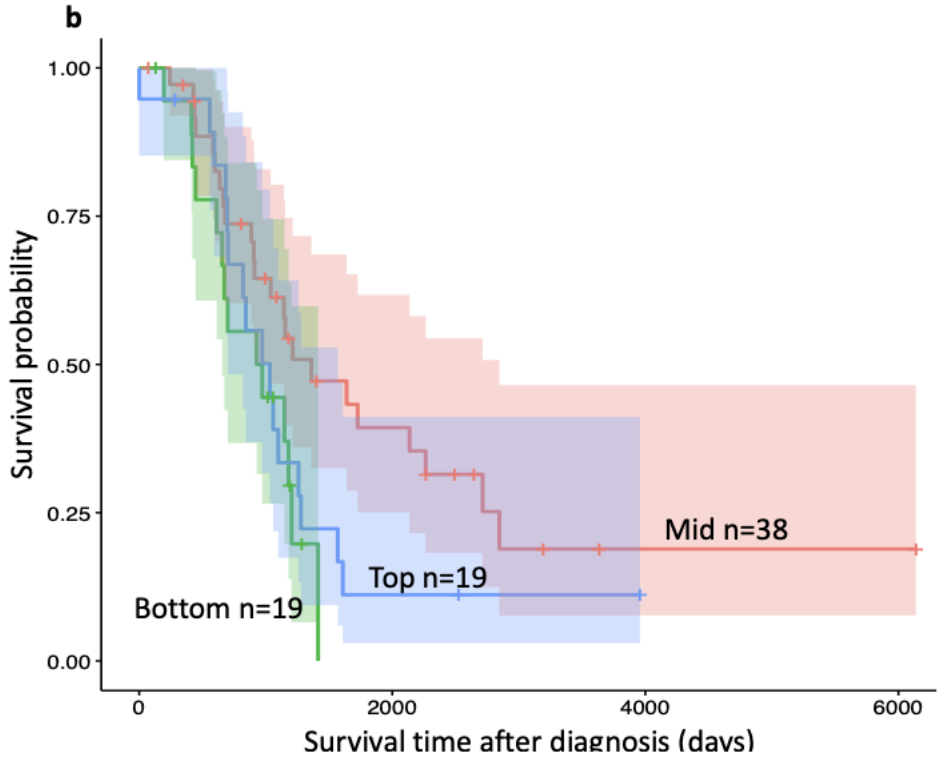
c

Condition		Hazard ratio	P value	Forest plot
Age	N=155	1.01 (0.99-1.03)	0.479	
Stage	N=155	1.11 (0.73-1.68)	0.627	
HRD	Absent N=57	Reference		
	Present N=98	0.38 (0.24-0.61)	0.001	
Cohort	SHGSOC N=50	Reference		
	AOCS N=46	2.18 (1.24-3.82)	0.006	
	BCCA N=25	0.82 (0.44-1.53)	0.542	
	TCGA N=15	2.40 (1.23-4.66)	0.001	
Chromoplexy Severity	Middle N=59	Reference		
	Top N=42	0.69 (0.41-1.17)	0.169	
	Bottom N=54	0.90 (0.54-1.49)	0.169	

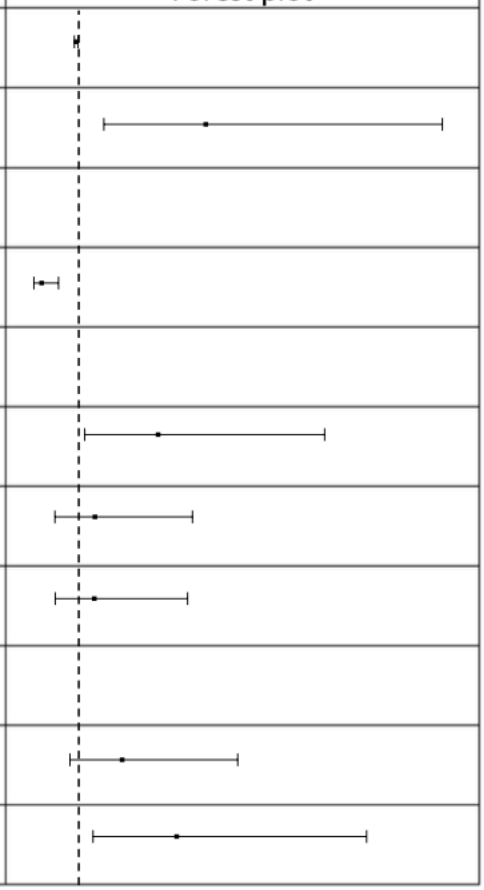
Figure 61 Chromoplexy severity and survival

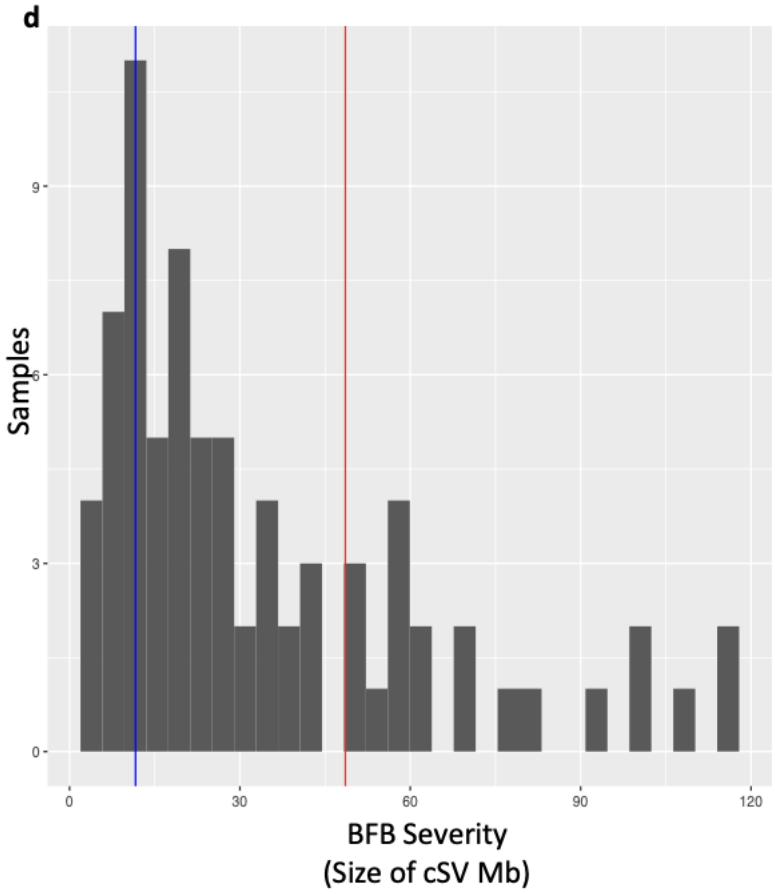
The distribution of severity of chromoplexy measured by number of SV Explained by chromoplexy in samples with chromoplexy. **a** The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups the highest genomic instability (green), the least genomic instability (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by chromoplexy (top 25%) to the samples least effected by chromoplexy (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.

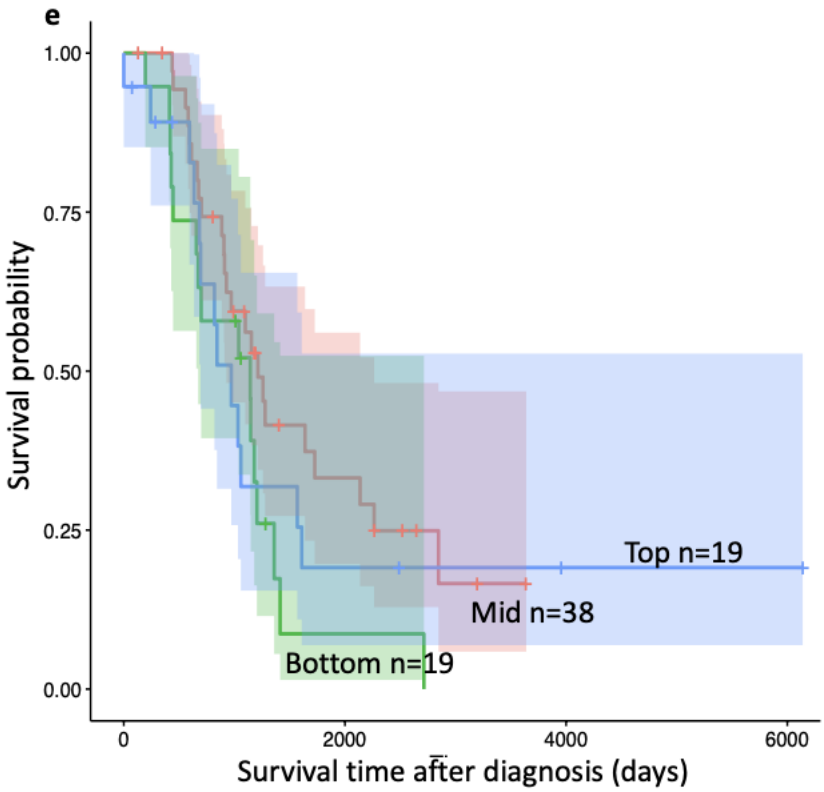

a
Samples
10
5
0
0
50
100
150
BFB Severity
(Number of SVs)


b
Survival probability
1.00
0.75
0.50
0.25
0.00
Mid n=38
Top n=19
Bottom n=19
0
2000
4000
6000
Survival time after diagnosis (days)

c

Condition		Hazard ratio	P value	Forest plot
Age	N=76	1.00 (0.97-1.05)	0.628	
Stage	N=76	3.7 (1.58-8.57)	0.003	
HRD	Absent N=55	Reference		
	Present N=21	0.30 (0.14-0.65)	0.002	
Cohort	SHGSOC N=25	Reference		
	AOCS N=21	2.70 (1.19-6.14)	0.017	
	BCCA N=12	1.40 (0.57-3.41)	0.461	
	TCGA N=9	1.40 (0.58-3.31)	0.465	
BFB Severity	Middle N=38	Reference		
	Top N=19	2.00 (0.88-4.34)	0.099	
	Bottom N=19	3.10 (1.36-6.99)	0.007	


d
Samples
9
6
3
0
0
30
60
90
120
BFB Severity
(Size of cSV Mb)

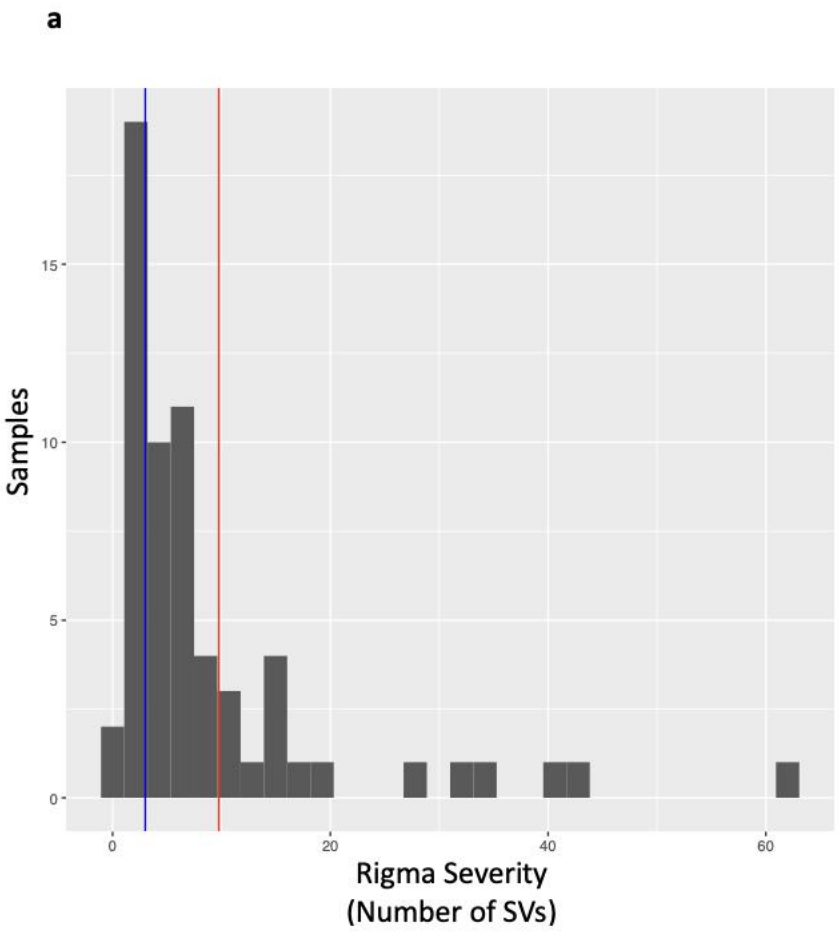

e
1.00
0.75
0.50
0.25
0.00
Survival probability
Top n=19
Mid n=38
Bottom n=19
0
2000
4000
6000
Survival time after diagnosis (days)

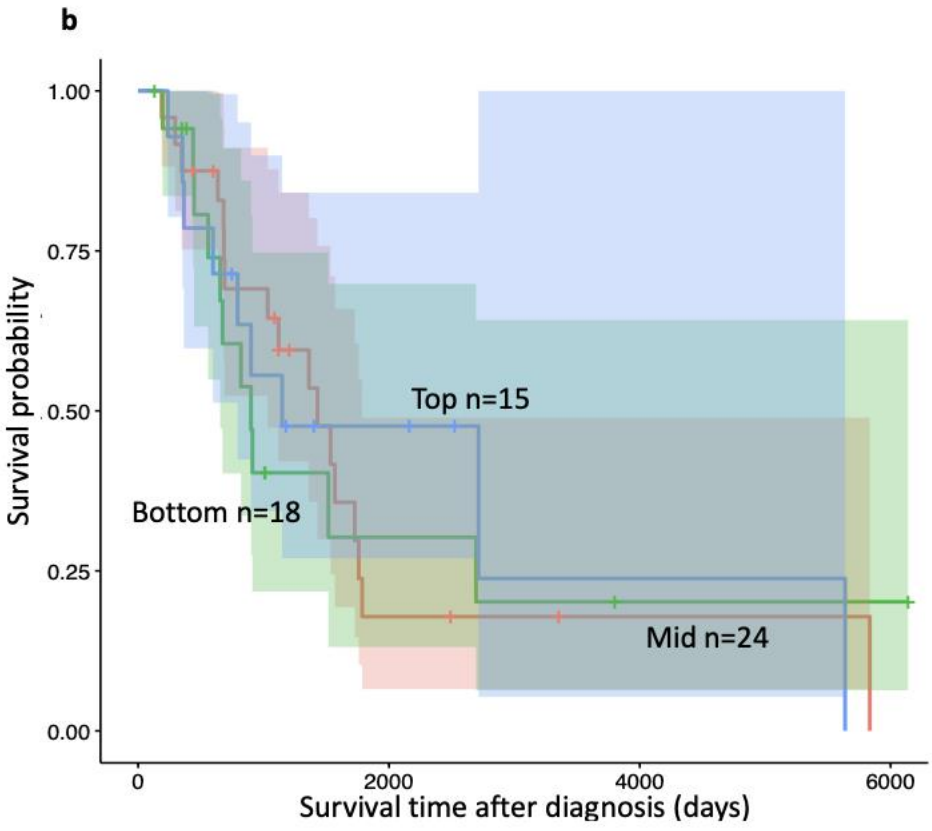
f

Condition		Hazard ratio	P value	Forest plot
Age	N=76	1.00 (0.98-1.06)	0.404	
Stage	N=76	3.25 (1.31-8.05)	0.011	
HRD	Absent N=55	Reference		
	Present N=21	0.24 (0.11-0.54)	0.001	
Cohort	SHGSOC N=25	Reference		
	AOCS N=21	2.68 (1.14-6.29)	0.024	
	BCCA N=12	1.72 (0.68-4.37)	0.256	
	TCGA N=9	1.86 (0.78-4.44)	0.16	
BFB Severity	Middle N=38	Reference		
	Top N=19	1.72 (1.81-3.67)	0.161	
	Bottom N=19	2.33 (1.07-5.06)	0.032	

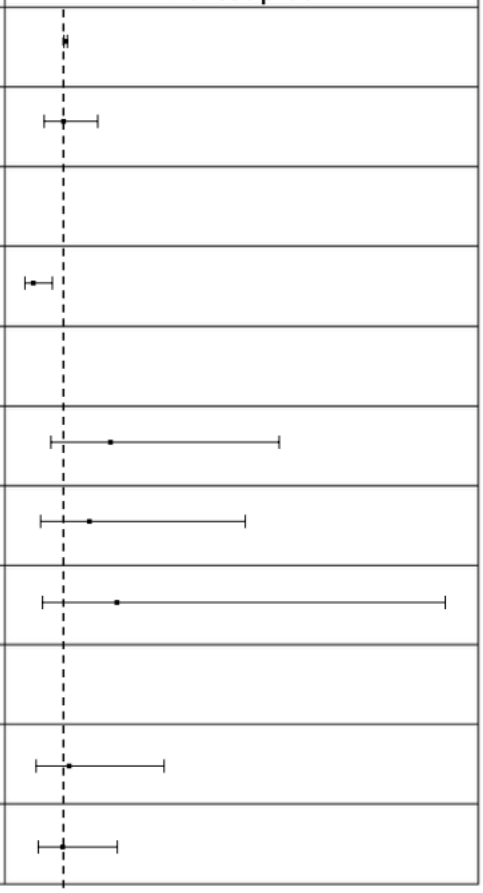
Figure 62 The least severe BFB show increased risk of death

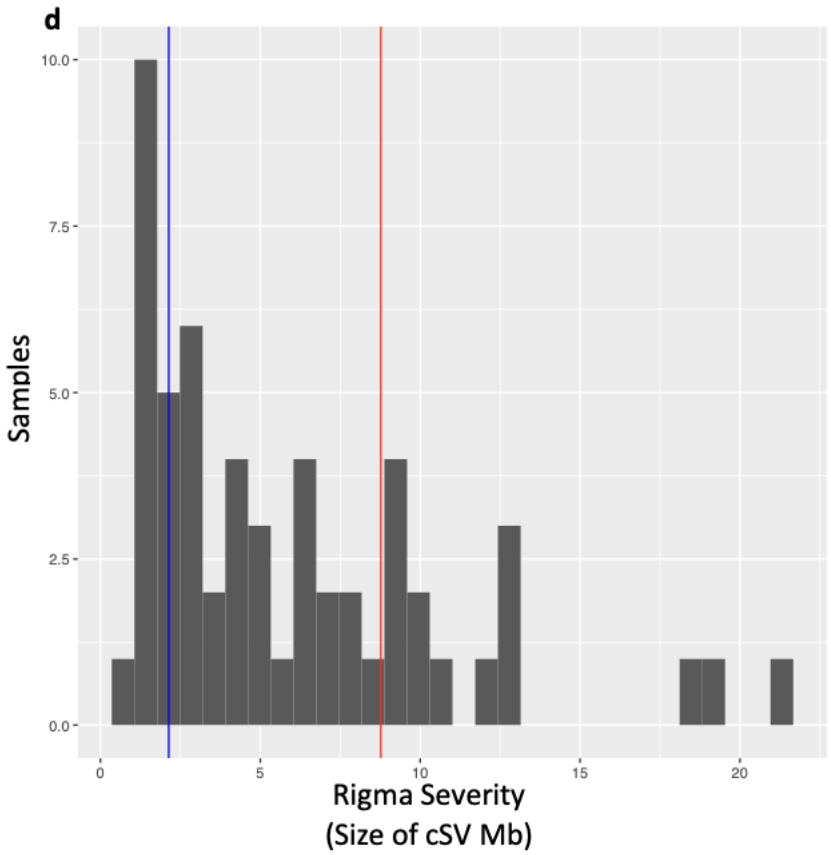
The distribution of severity of BFB measured By number of SV explained by BFB in samples with BFB. **a** The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by BFB (top 25%) to the samples least effected by BFB (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of BFB measured by Mb covered by BFB in samples with BFB. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **f** A Cox proportional hazards model comparing the survival of the samples most severely affected by BFB (top 25%) to the samples least effected by BFB (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.

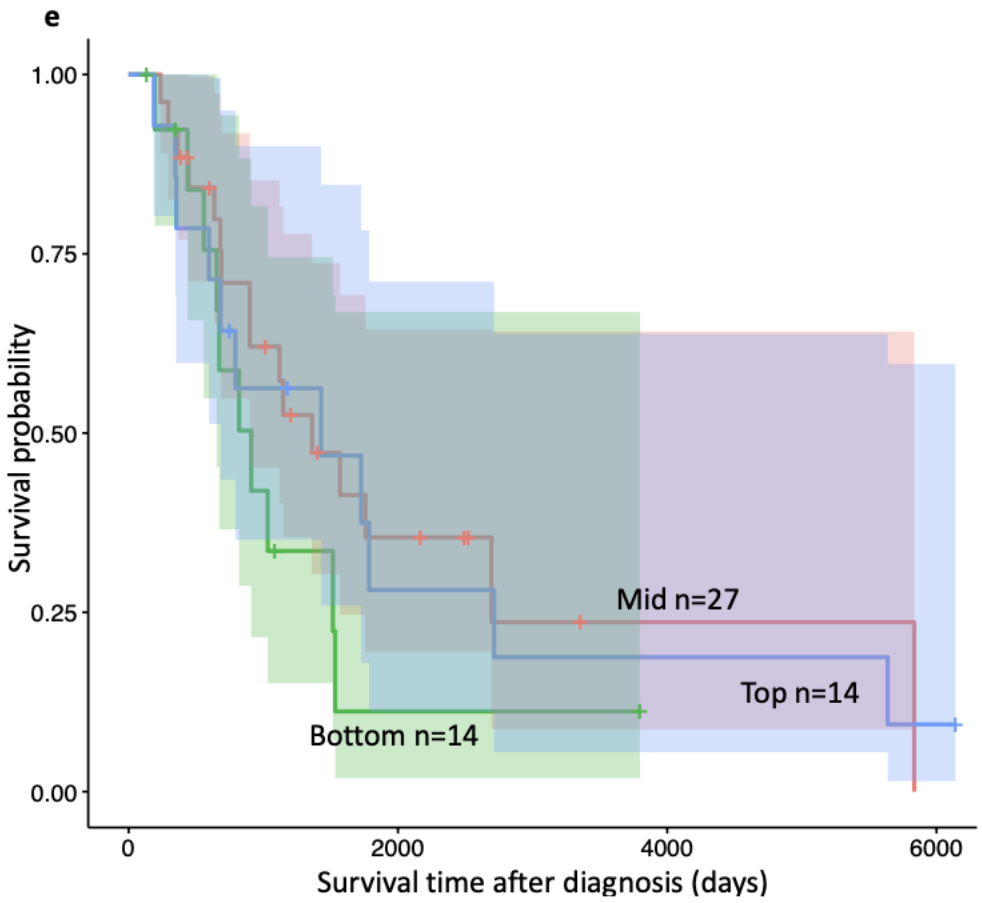

a
Samples
15
10
5
0
0
20
40
60
Rigma Severity
(Number of SVs)


b
Survival probability
1.00
0.75
0.50
0.25
0.00
Top n=15
Bottom n=18
Mid n=24
0
2000
4000
6000
Survival time after diagnosis (days)

c

Condition		Hazard ratio	P value	Forest plot
Age	N=57	1.01 (0.98-1.05)	0.455	
Stage	N=57	1.02 (1.58-1.80)	0.944	
HRD	Absent N=31	Reference		
	Present N=26	0.33 (0.15-0.77)	0.01	
Cohort	SHGSOC N=23	Reference		
	AOCS N=13	2.09 (0.74-5.92)	1.165	
	BCCA N=10	1.61 (0.50-5.15)	0.422	
	TCGA N=3	2.18 (0.49-9.65)	0.304	
Rigma Severity	Middle N=24	Reference		
	Top N=15	1.15 (0.40-3.30)	0.998	
	Bottom N=18	1.00 (0.45-2.24)	0.80	


d
10.0
7.5
5.0
2.5
0.0
Samples
0
5
10
15
20
Rigma Severity
(Size of cSV Mb)

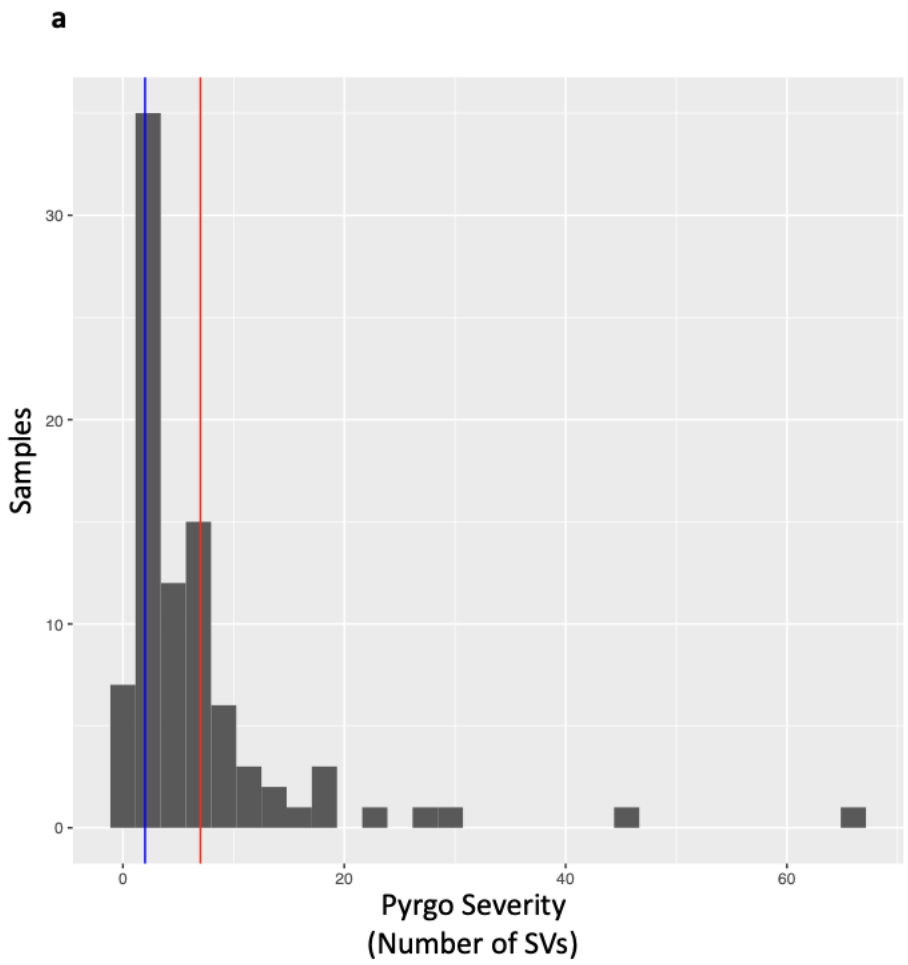

e
1.00
0.75
0.50
0.25
0.00
Survival probability
Mid n=27
Top n=14
Bottom n=14
0
2000
4000
6000
Survival time after diagnosis (days)

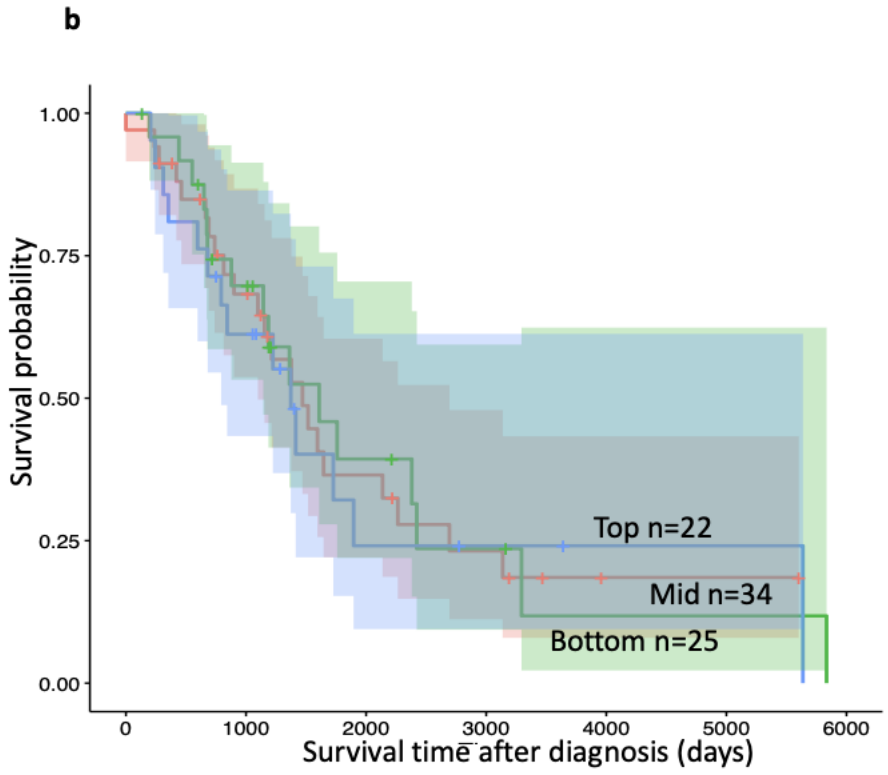
f

Condition		Hazard ratio	P value	Forest plot
Age	N=55	1.02 (0.98-1.06)	0.378	
Stage	N=55	1.06 (0.61-1.86)	0.835	
HRD	Absent N=31	Reference		
	Present N=24	0.25 (0.08-0.72)	0.001	
Cohort	SHGSOC N=23	Reference		
	AOCS N=13	1.79 (0.66-4.66)	0.235	
	BCCA N=10	1.73 (0.54-5.56)	0.357	
	TCGA N=3	2.03 (0.51-8.10)	0.316	
Rigma Severity	Middle N=27	Reference		
	Top N=14	1.79 (0.65-4.96)	0.808	
	Bottom N=14	2.33 (0.47-2.62)	0.260	

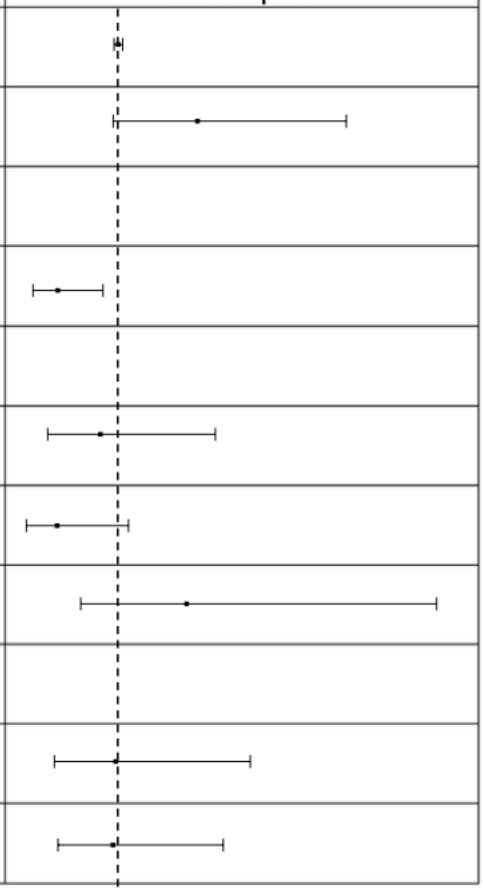
Figure 63 Rigma severity and survival

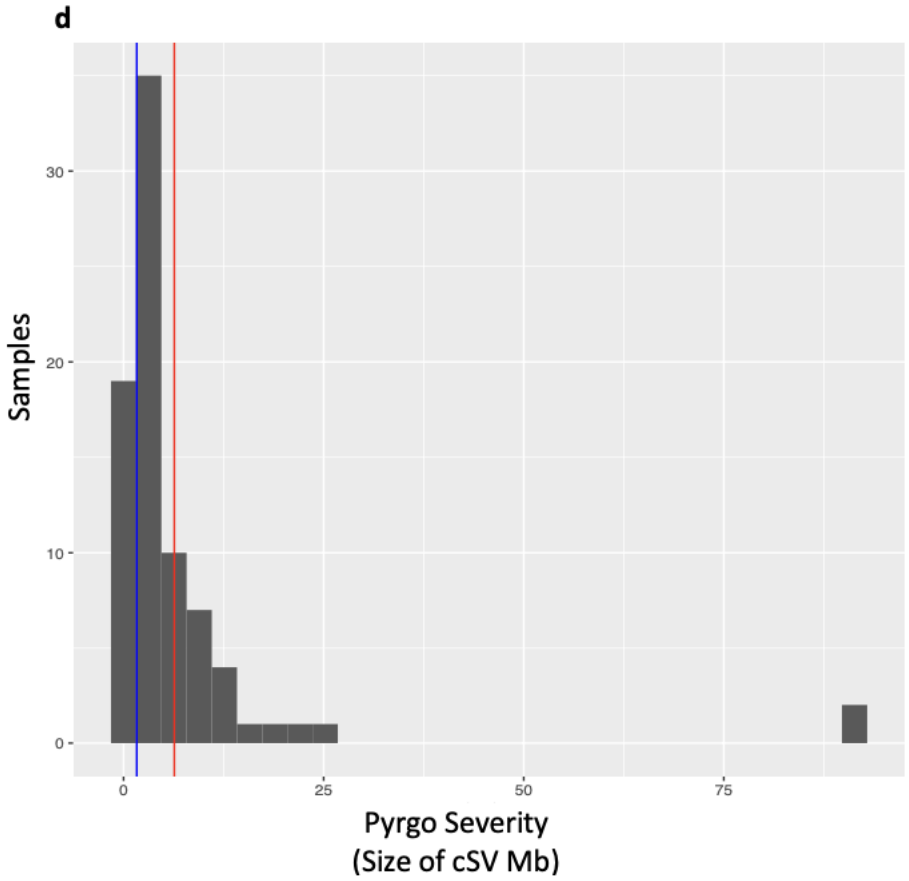
The distribution of severity of rigma measured by number of SV explained by rigma in samples with rigma. **a** The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red) **c**. A Cox proportional hazards model comparing the survival of the samples most severely affected by rigma (top 25%) to the samples least effected by rigma (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of rigma measured by Mb covered by rigma in samples with rigma. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **f** A Cox proportional hazards model comparing the survival of the samples most severely affected by rigma (top 25%) to the samples least effected by rigma (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.

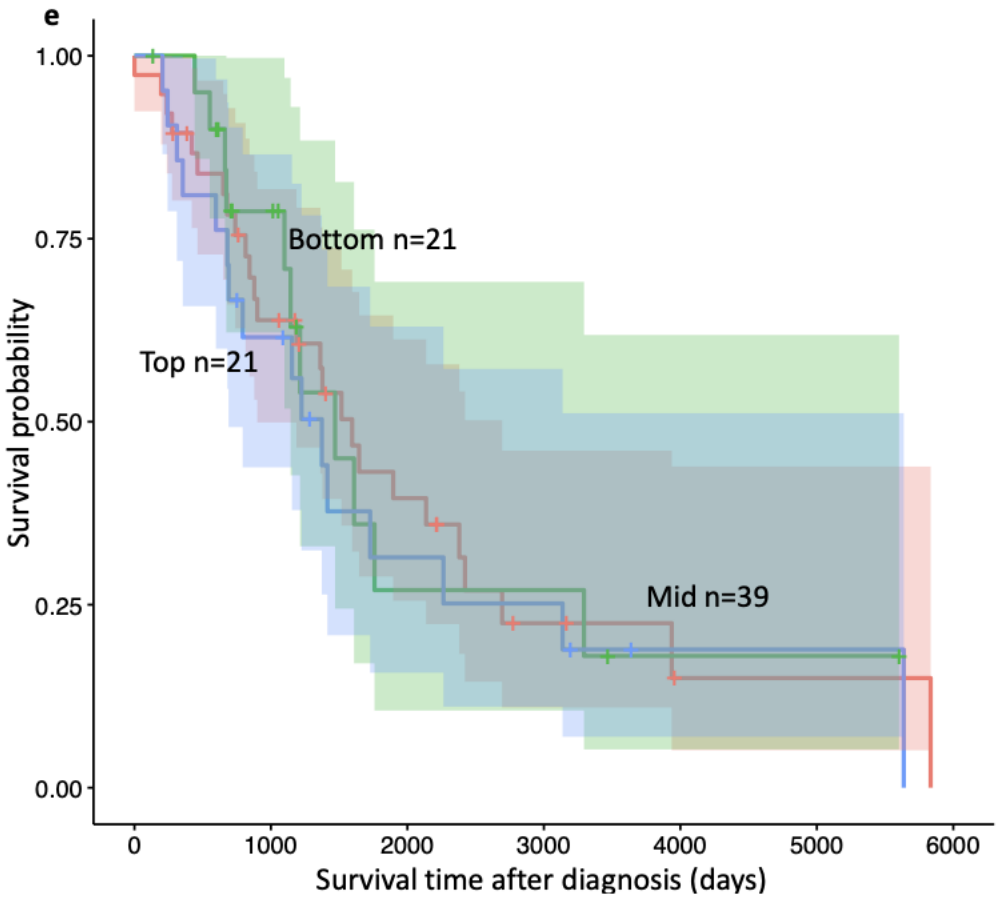

a
Samples
30
20
10
0
0
20
40
60
Pyrgo Severity
(Number of SVs)


b
Survival probability
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
6000
Survival time after diagnosis (days)
Top n=22
Mid n=34
Bottom n=25

c

Condition		Hazard ratio	P value	Forest plot
Age	N=81	1.01 (0.98-1.05)	0.496	
Stage	N=81	1.71 (0.97-3.02)	0.065	
HRD	Absent N=41	Reference		
	Present N=40	0.48 (0.26-0.88)	0.017	
Cohort	SHGSOC N=31	Reference		
	AOCS N=16	0.89 (0.39-1.87)	0.695	
	BCCA N=16	0.47 (0.20-1.15)	0.083	
	TCGA N=8	1.61 (0.66-3.82)	0.276	
Pyrgo Severity	Middle N=34	Reference		
	Top N=25	0.99 (0.45-2.17)	0.960	
	Bottom N=22	0.97 (0.46-1.94)	0.923	


d
Samples
30
20
10
0
0
25
50
75
Pyrgo Severity
(Size of cSV Mb)


e
1.00
0.75
0.50
0.25
0.00
Survival probability
Bottom n=21
Top n=21
Mid n=39
0
1000
2000
3000
4000
5000
6000
Survival time after diagnosis (days)

f

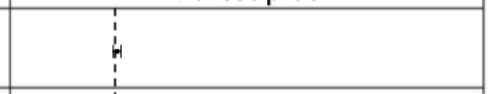
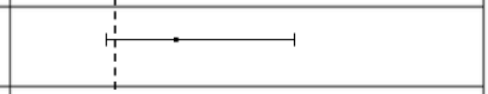
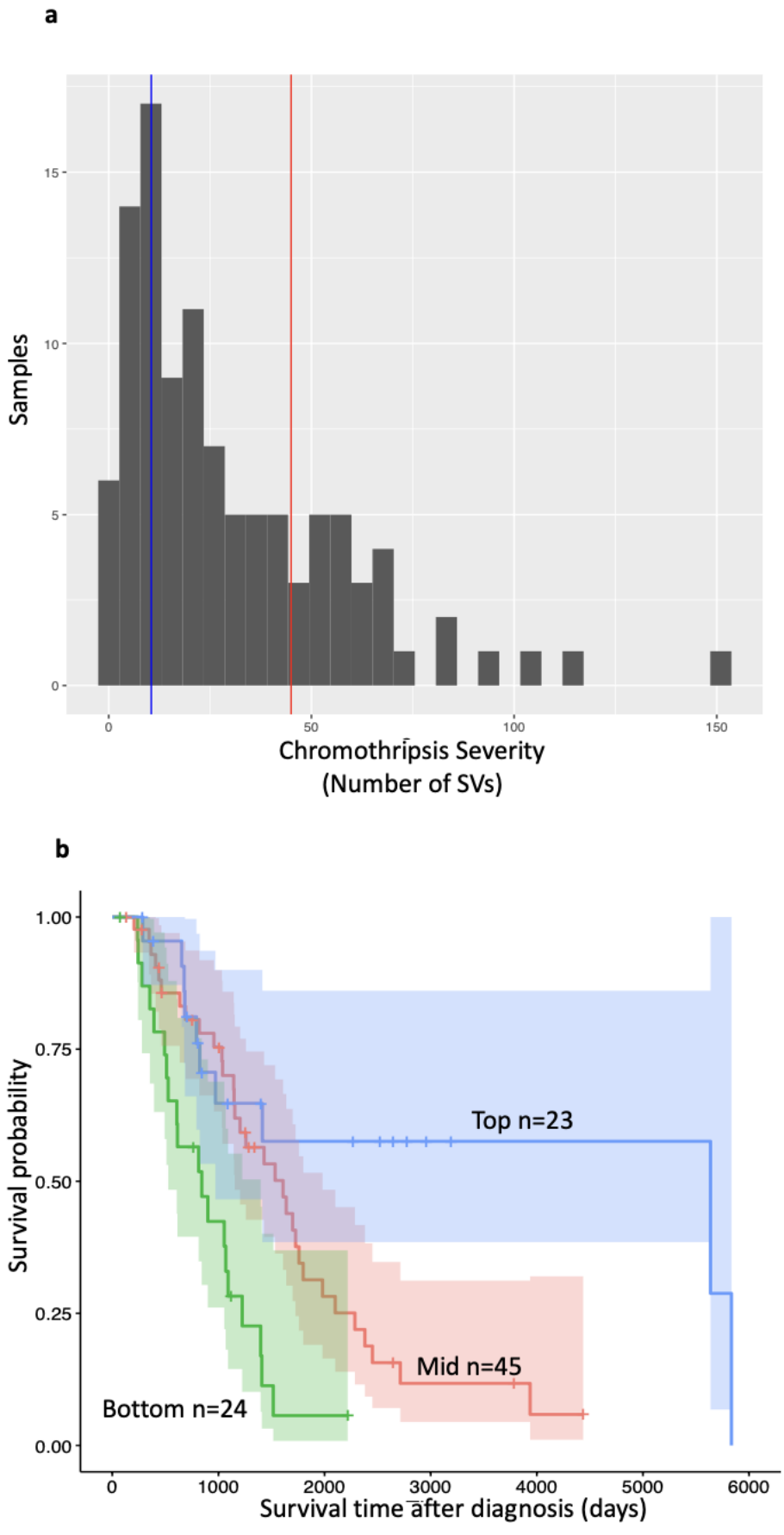
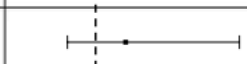
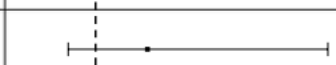
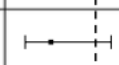
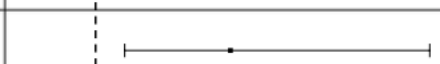
Condition		Hazard ratio	P value	Forest plot
Age	N=81	1.01 (0.98-1.05)	0.454	
Stage	N=81	1.55 (0.91-2.63)	0.104	
HRD	Absent N=41	Reference		
	Present N=40	0.45 (0.25-0.83)	0.011	
Cohort	SHGSOC N=31	Reference		
	AOCS N=16	0.93 (0.43-2.04)	0.86	
	BCCA N=17	0.49 (0.22-1.09)	0.082	
	TCGA N=8	1.71 (0.73-3.98)	0.216	
Pyrgo Severity	Middle N=39	Reference		
	Top N=21	0.85 (0.42-1.70)	0.637	
	Bottom N=11	0.85 (0.39-1.70)	0.593	

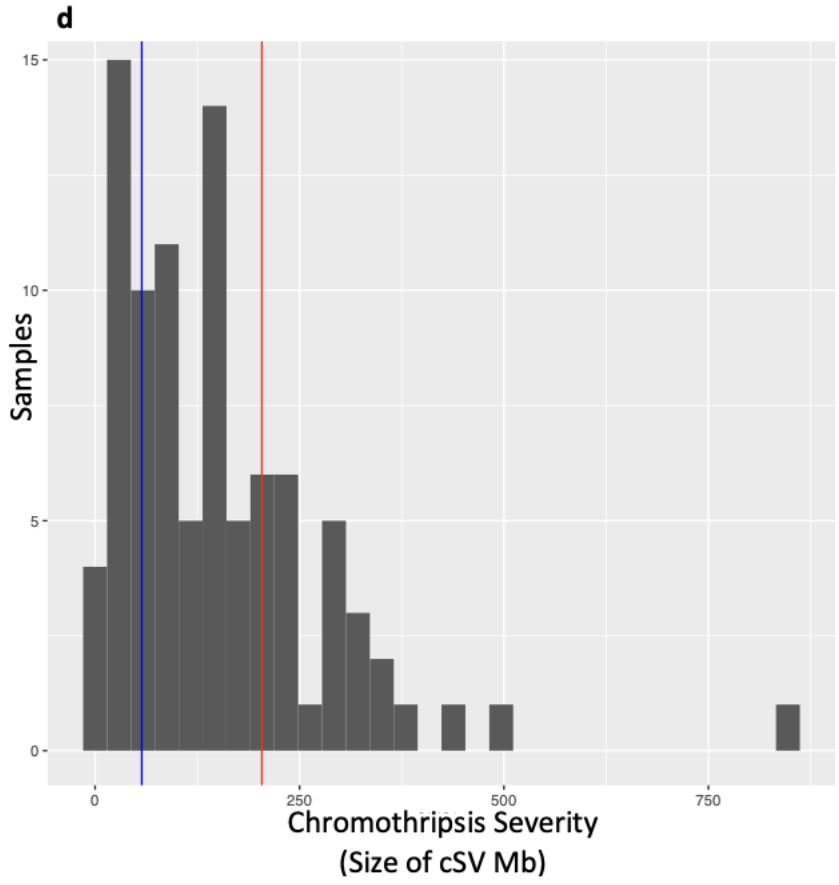
Figure 64 Pyrgo severity and survival

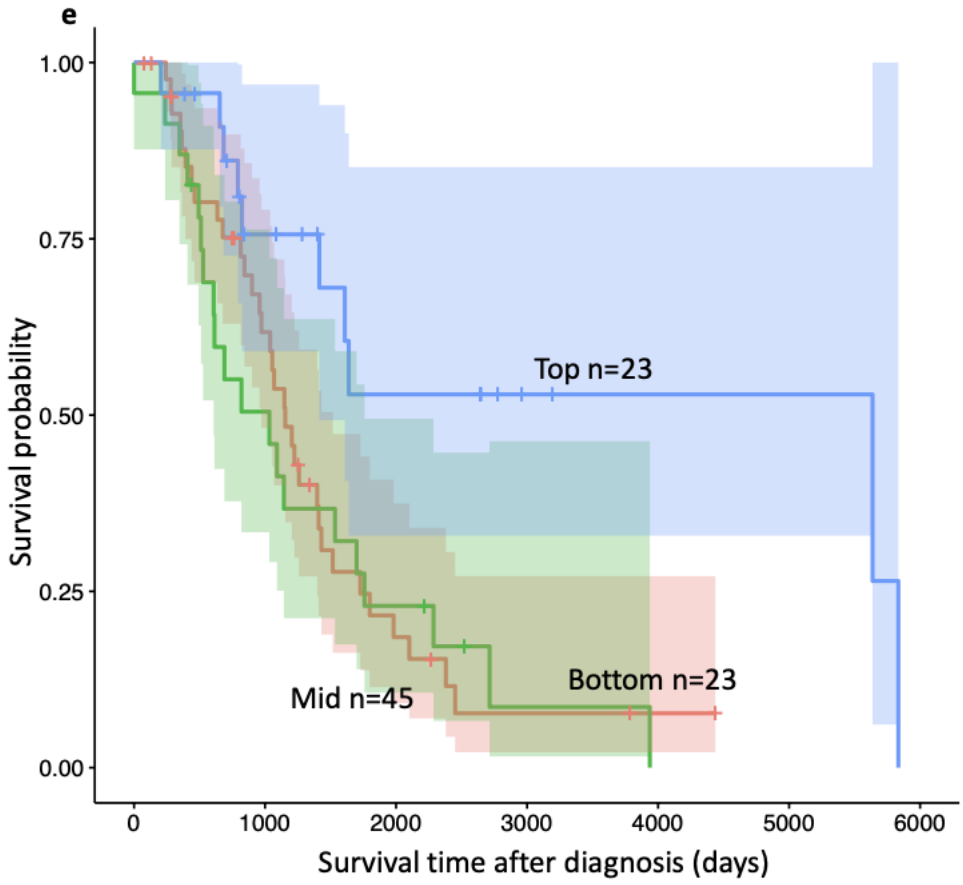
a The distribution of severity of pyrgo measured by number of SV explained by pyrgo in samples with pyrgo. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by pyrgo (top 25%) to the samples least effected by pyrgo (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of pyrgo measured by Mb covered by pyrgo in samples with pyrgo. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **f** A Cox proportional hazards model comparing the survival of the samples most severely affected by pyrgo (top 25%) to the samples least effected by pyrgo (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.


a
Samples
15
10
5
0
0
50
100
150
Chromothripsis Severity
(Number of SVs)
b
Survival probability
1.00
0.75
0.50
0.25
0.00
Top n=23
Mid n=45
Bottom n=24
0
1000
2000
3000
4000
5000
6000
Survival time after diagnosis (days)

c

Condition		Hazard ratio	P value	Forest plot
Age	N=92	1.03 (1.00-1.07)	0.053	
Stage	N=92	1.20 (0.72-2.01)	0.482	
HRD	Absent N=43	Reference		
	Present N=49	0.51 (0.29-0.91)	0.022	
Cohort	SHGSOC N=31	Reference		
	AOCS N=34	1,31 (0.67-2.55)	0.431	
	BCCA N=6	1.43 (0.42-4.88)	0.572	
	TCGA N=10	1.55 (0.68-3.52)	0.299	
Chromothripsis Severity	Middle N=45	Reference		
	Top N=23	0.49 (0.21-2.15)	0.101	
	Bottom N=24	2.45 (1.30-4.64)	0.006	


d
15
10
5
0
Samples
0
250
500
750
Chromothripsis Severity
(Size of cSV Mb)

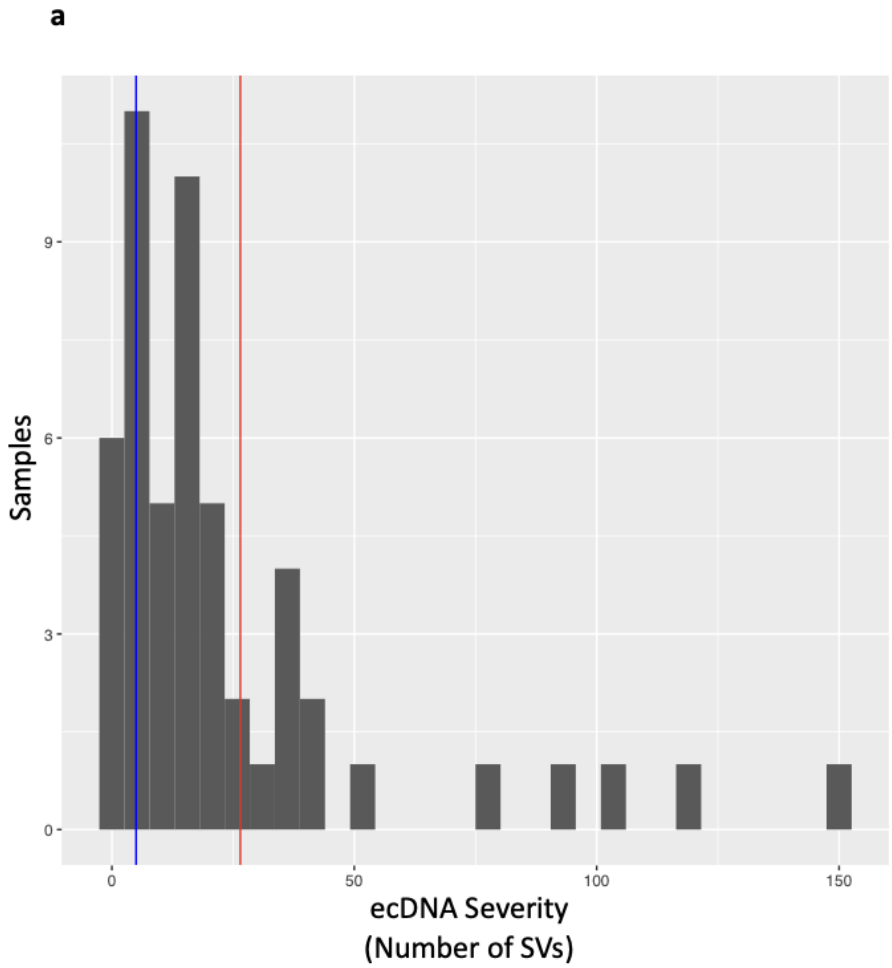

e
1.00
0.75
0.50
0.25
0.00
Survival probability
Top n=23
Bottom n=23
Mid n=45
0
1000
2000
3000
4000
5000
6000
Survival time after diagnosis (days)

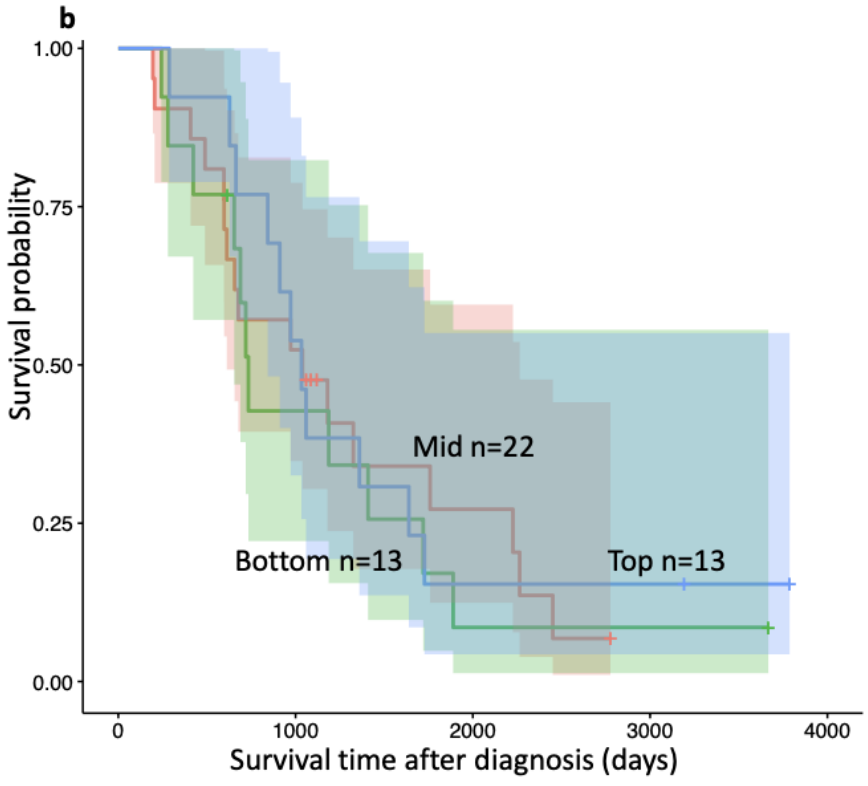
f

Condition		Hazard ratio	P value	Forest plot
Age	N=91	1.02 (0.99-1.06)	0.164	
Stage	N=91	1.21 (0.74-1.99)	0.448	
HRD	Absent N=44	Reference		
	Present N=47	0.42 (0.23-0.76)	0.005	
Cohort	SHGSOC N=31	Reference		
	AOCS N=34	1.54 (0.76-3.13)	0.232	
	BCCA N=6	1.15 (0.35-3.80)	0.816	
	TCGA N=10	1.55 (0.68-1.49)	0.297	
Chromothripsis Severity	Middle N=45	Reference		
	Top N=23	0.31 (0.14-0.72)	0.007	
	Bottom N=23	0.78 (0.40-1.49)	0.446	

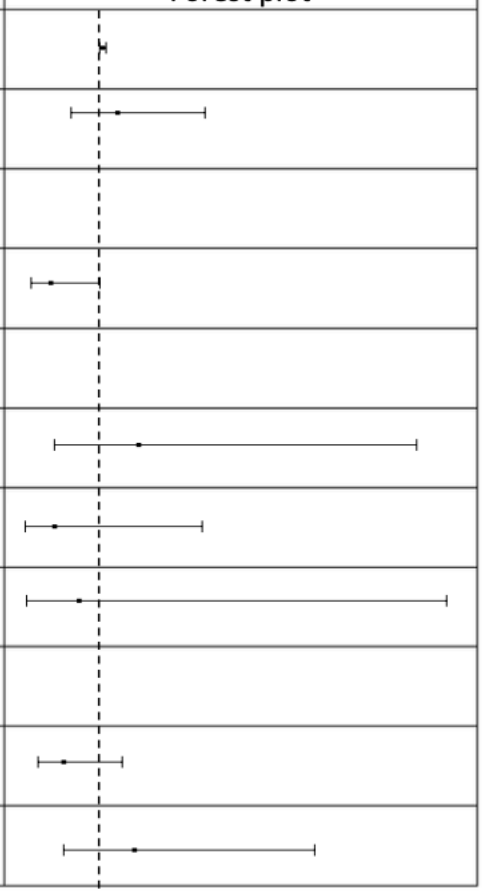
Figure 65 Increased chromothripsis severity show a decrease in risk of death

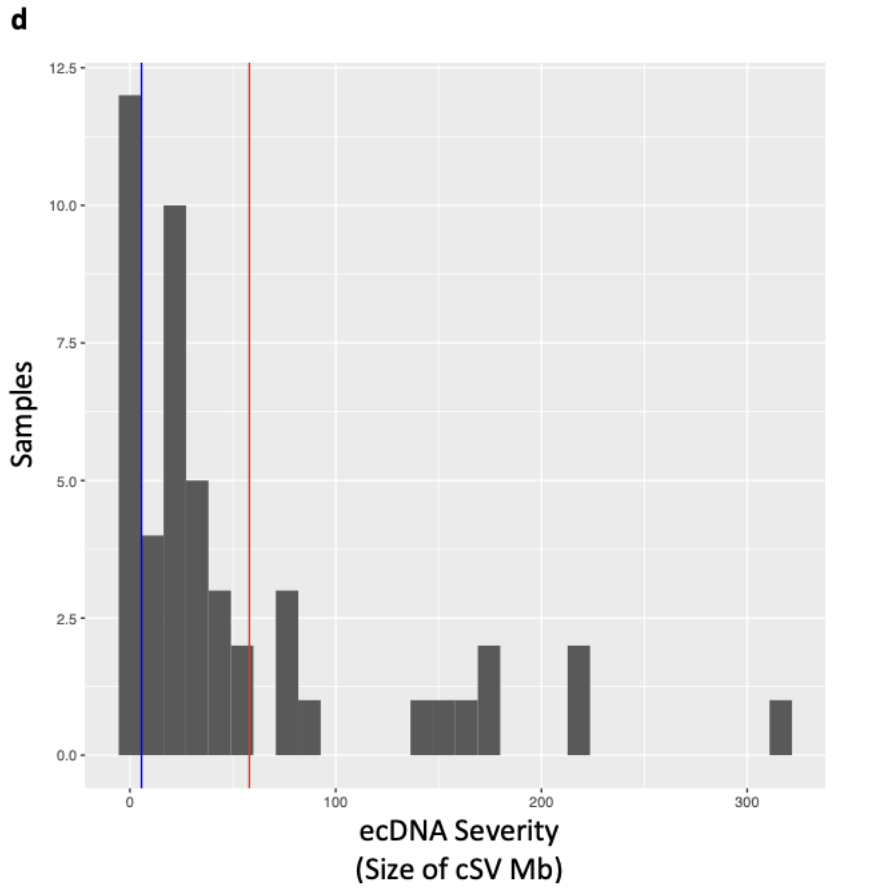
a The distribution of severity of chromothripsis measured by number of SV explained by chromothripsis in samples with chromothripsis. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by chromothripsis (top 25%) to the samples least effected by chromothripsis (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of chromothripsis measured by Mb covered by chromothripsis in samples with chromothripsis.The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **f** A Cox proportional hazards model comparing the survival of the samples most severely affected by chromothripsis (top 25%) to the samples least effected by chromothripsis (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.

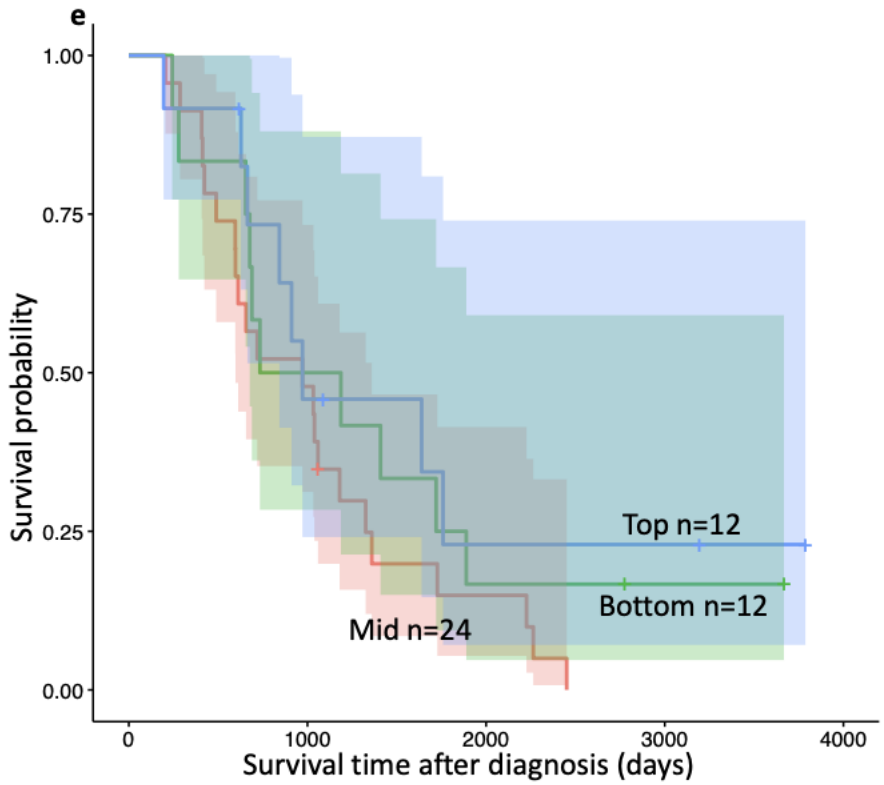

a
Samples
9
6
3
0
0
50
100
150
ecDNA Severity
(Number of SVs)


b
Survival probability
1.00
0.75
0.50
0.25
0.00
Mid n=22
Bottom n=13
Top n=13
0
1000
2000
3000
4000
Survival time after diagnosis (days)

c

Condition		Hazard ratio	P value	Forest plot
Age	N=48	1.02 (0.98-1.06)	0.359	
Stage	N=48	1.20 (0.64-2.24)	0.577	
HRD	Absent N=27	Reference		
	Present N=21	0.40 (0.16-0.98)	0.046	
Cohort	SHGSOC N=7	Reference		
	AOCS N=33	1.45 (0.44-4.79)	0.544	
	BCCA N=3	0.44 (0.088-2.21)	0.319	
	TCGA N=2	0.73 (0.11-5.15)	0.756	
ecDNA Severity	Middle N=22	Reference		
	Top N=13	0.55 (0.24-1.25)	0.482	
	Bottom N=13	1.40 (0.55-3.56)	0.154	


d
12.5
10.0
7.5
5.0
2.5
0.0
Samples
0
100
200
300
ecDNA Severity
(Size of cSV Mb)

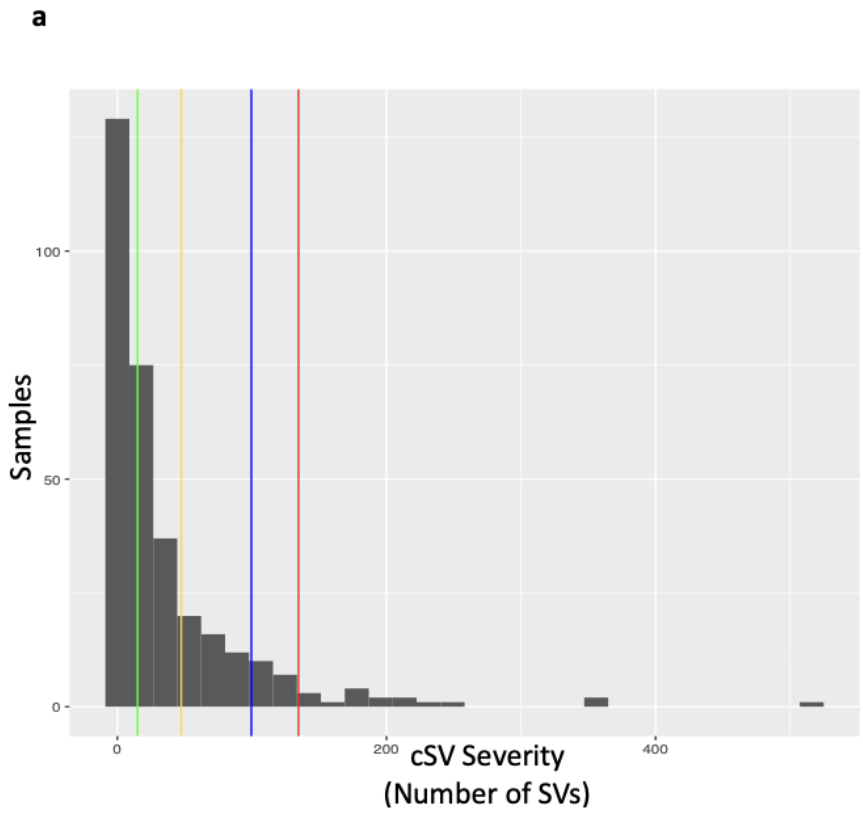

e
1.00
0.75
0.50
0.25
0.00
Survival probability
Top n=12
Bottom n=12
Mid n=24
0
1000
2000
3000
4000
Survival time after diagnosis (days)

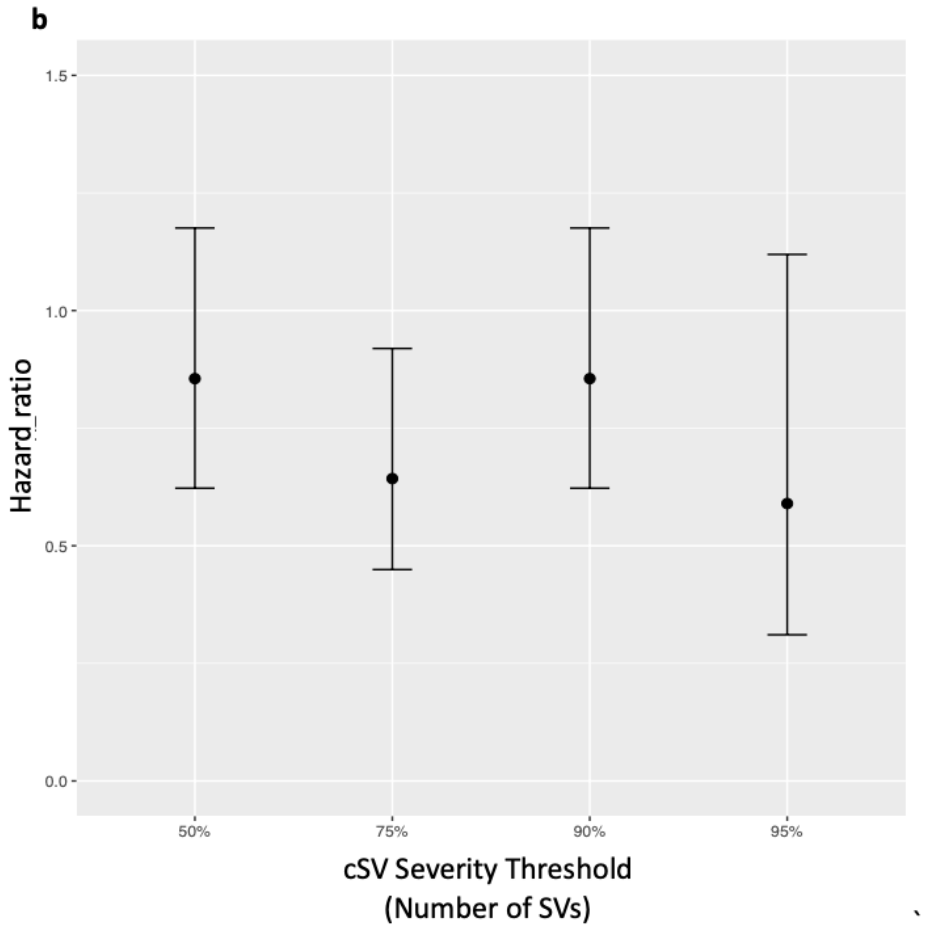
f

Condition		Hazard ratio	P value	Forest plot
Age	N=48	1.03 (0.99-1.07)	0.13	
Stage	N=48	1.07 (0.47-2.42)	0.879	
HRD	Absent N=27	Reference		
	Present N=21	0.22 (0.08-0.61)	0.004	
Cohort	SHGSOC N=7	Reference		
	AOCS N=33	6.56 (1.69-24.49)	0.007	
	BCCA N=3	0.55 (0.12-2.25)	0.438	
	TCGA N=2	2.27 (0.34-15.12)	0.396	
ecDNA Severity	Middle N=24	Reference		
	Top N=12	0.13 (0.04-0.42)	0.001	
	Bottom N=12	0.44 (0.18-1.06)	0.067	

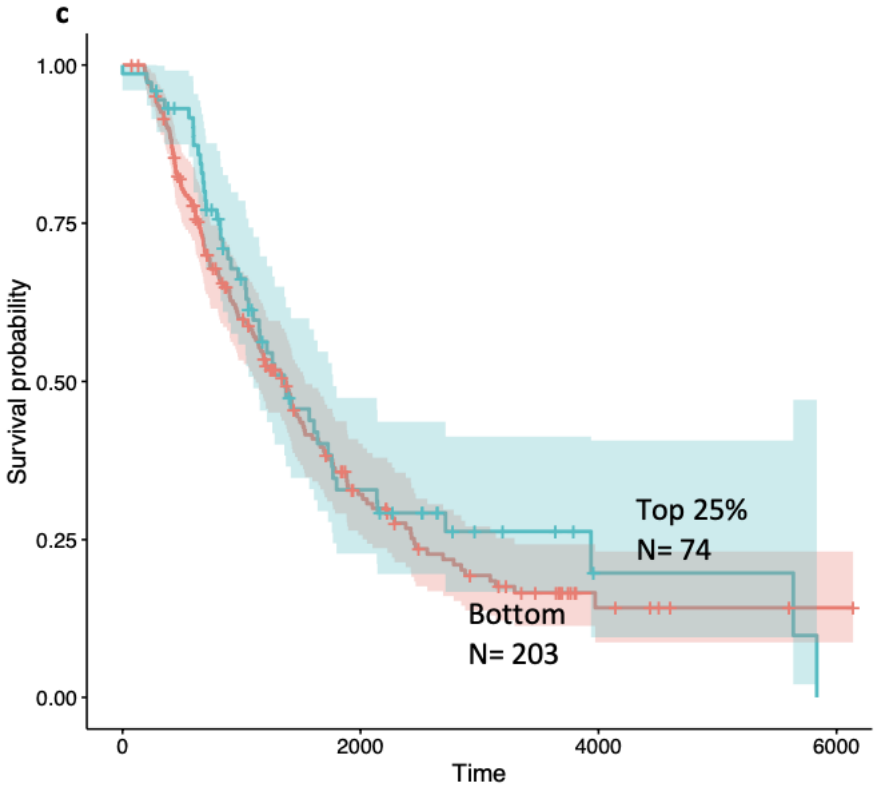
Figure 66 ecDNA severity and survival

a The distribution of severity of ecDNA measured by number of SV explained by ecDNA in samples with ecDNA. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by ecDNA (top 25%) to the samples least effected by ecDNA (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of ecDNA measured by Mb covered by ecDNA in samples with ecDNA. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **f** A Cox proportional hazards model comparing the survival of the samples most severely affected by ecDNA (top 25%) to the samples least effected by ecDNA (bottom 25%) adjusting for age, stage at diagnosis, HRD status d sub-cohort.


a
Samples
100
50
0
0
200
400
cSV Severity
(Number of SVs)

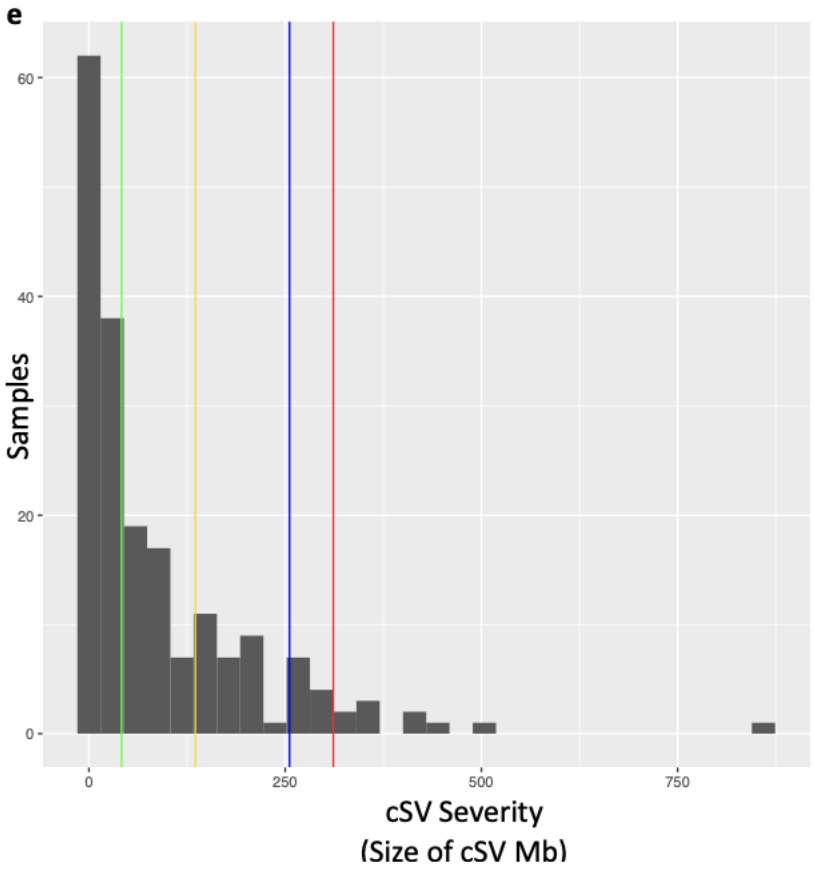

b
Hazard_ratio
1.5
1.0
0.5
0.0
50%
75%
90%
95%
cSV Severity Threshold
(Number of SVs)

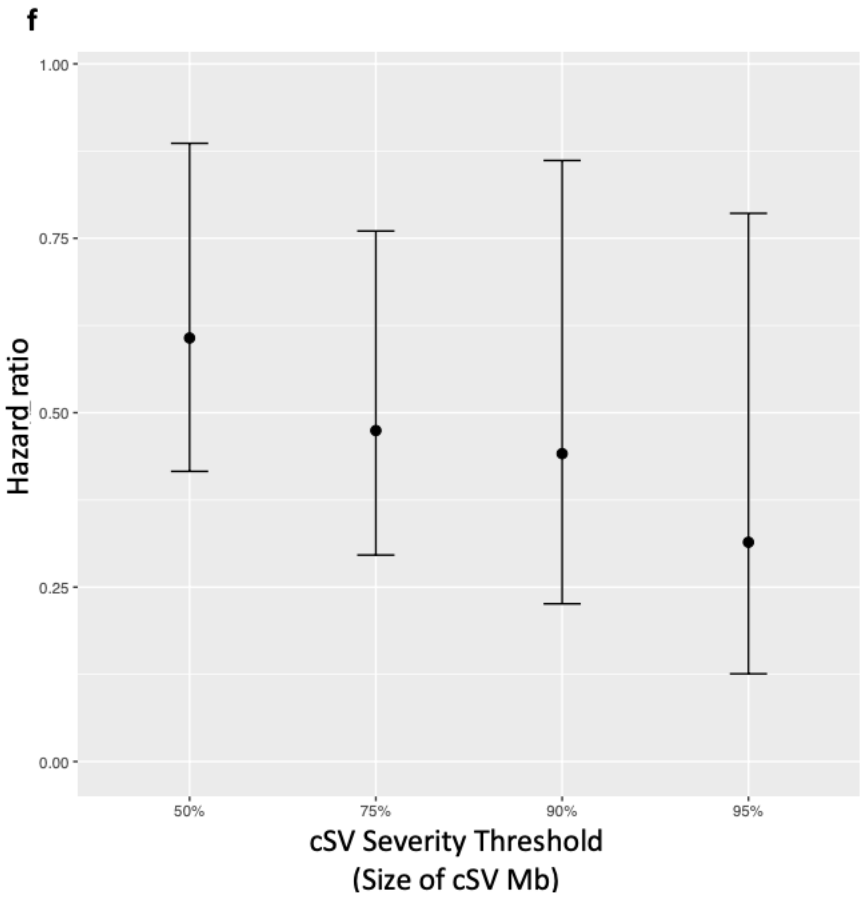
c

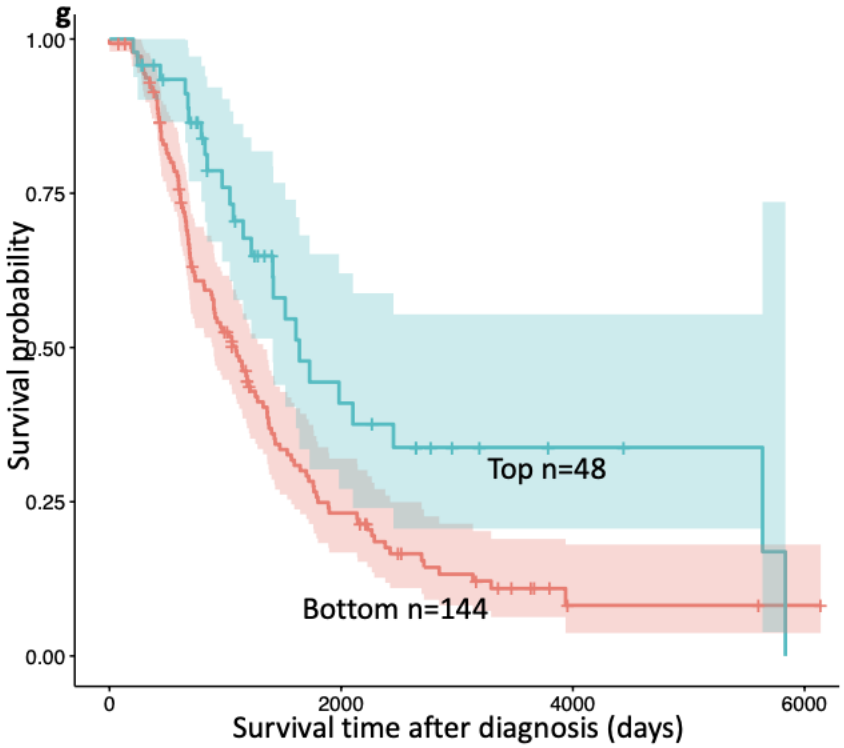


d

Condition		Hazard ratio	P value	Forest plot
Age	N=277	1.02 (1.00-1.04)	0.017	
Stage	N=277	1.50 (1.12-2.00)	0.006	
HRD	Absent N=118	Reference		
	Present N=159	0.44 (0.32-0.61)	0.001	
Cohort	SHGSOC N=81	Reference		
	AOCS N=80	1.77 (1.20-2.63)	0.004	
	BCCA N=59	0.79 (0.51-1.22)	0.286	
	TCGA N=31	1.55 (0.97-2.47)	0.065	
Severity of clusters	Bottom N=203	Reference		
	Top 25% N=74	0.64 (0.45-0.92)	0.016	


e
60
40
20
0
Samples
0
250
500
750
cSV Severity
(Size of cSV Mb)


f
1.00
0.75
0.50
0.25
0.00
Hazard ratio
50%
75%
90%
95%
cSV Severity Threshold
(Size of cSV Mb)

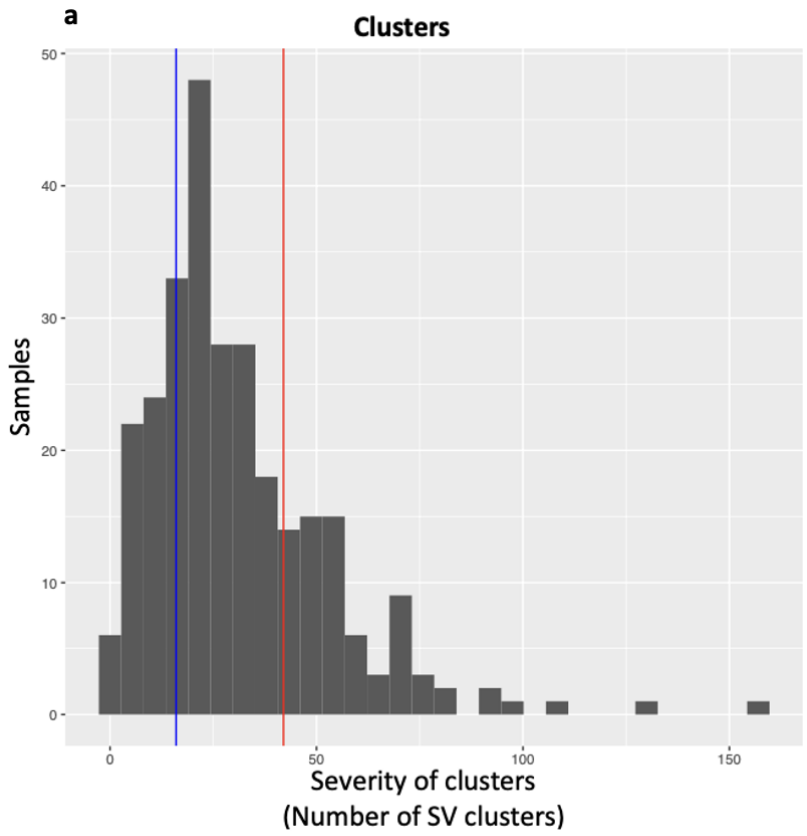


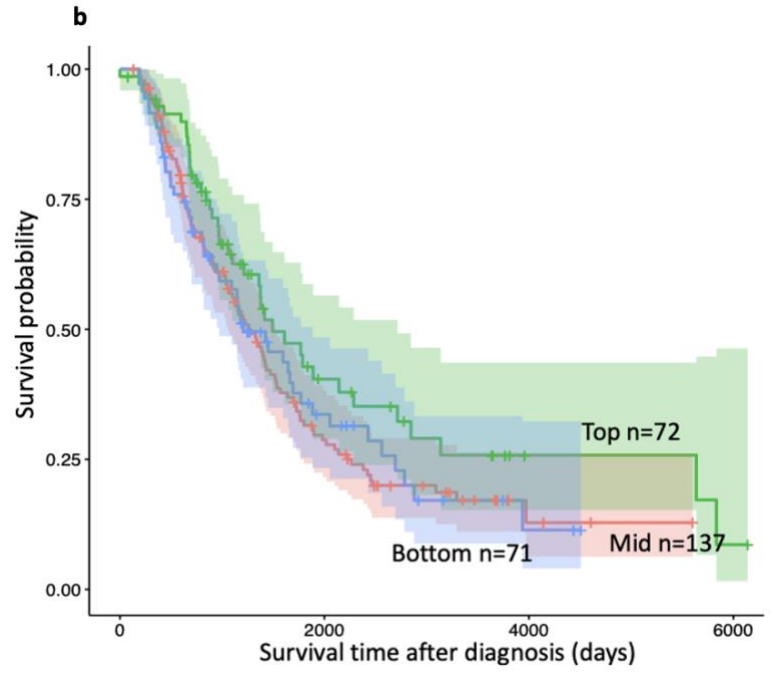
h

Condition		Hazard ratio	P value	Forest plot
Age	N=192	1.01 (0.99-1.03)	0.230	
Stage	N=192	1.40 (0.99-1.97)	0.054	
HRD	Absent N=95	Reference		
	Present N=97	0.50 (0.35-0.72)	0.001	
Cohort	SHGSOC N=58	Reference		
	AOCS N=59	1.41 (0.89-2.22)	0.143	
	BCCA N=33	0.59 (0.34-1.02)	0.056	
	TCGA N=22	1.28 (0.74-2.20)	0.378	
cSV Severity	Bottom N=144	Reference		
	Top N=48	0.47 (0.30-0.76)	0.002	

Figure 67 Increased total complex structural variant severity decrease risk of death

a The distribution of severity of all cSVs as measured by the total number of sv explained by any cSV in the sample with four lines show the quartile For 50% (green), 75% (yellow), 90% (blue), 95% (red). **b** The hazard ratio for a Cox proportional hazards model comparing the samples above and below the respective quartile (50%,75%,90%,95%) of genomically instable samples to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort. A hazard of 1 represent no effect and hazard ratios of less than 1 represent improved survival. **c** A Kaplan-Meier survival curve for time after diagnosis for two groups the most severely impacted 25% (blue), and remainder of the combined cohort (red). **d** A Cox proportional hazards model comparing the survival of the samples most severely affected by all cSVs (top 25%) to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort. **e** The distribution of severity of all cSVs as measured by the size of the genome impacted by any cSV in the sample with four lines show the quartile For 50% (green), 75% (yellow), 90% (blue), 95% (red). **f** The hazard ratio for a Cox proportional hazards model comparing the samples above and below the respective quartile (50%,75%,90%,95%) of genomically instable samples to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort. A hazard of 1 represent no effect and hazard ratios of less than 1 represent improved survival. **g** A Kaplan-Meier survival curve for time after diagnosis for two groups, the most severely impacted 50% (blue), and remainder of the combined cohort (red). **h** A Cox proportional hazards model comparing the survival of the samples most severely affected by all cSVs (top 25%) to the remainder of the combined cohort adjusting for age, stage at diagnosis, HRD status and sub-cohort.

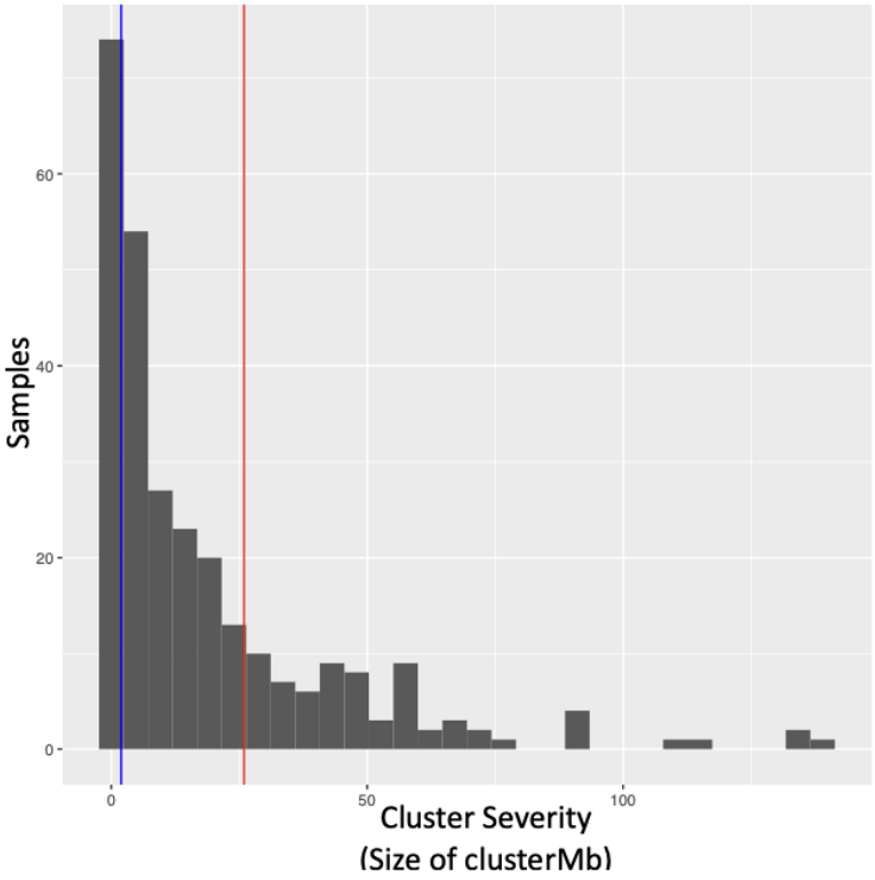

a
Clusters
Samples
Severity of clusters
(Number of SV clusters)

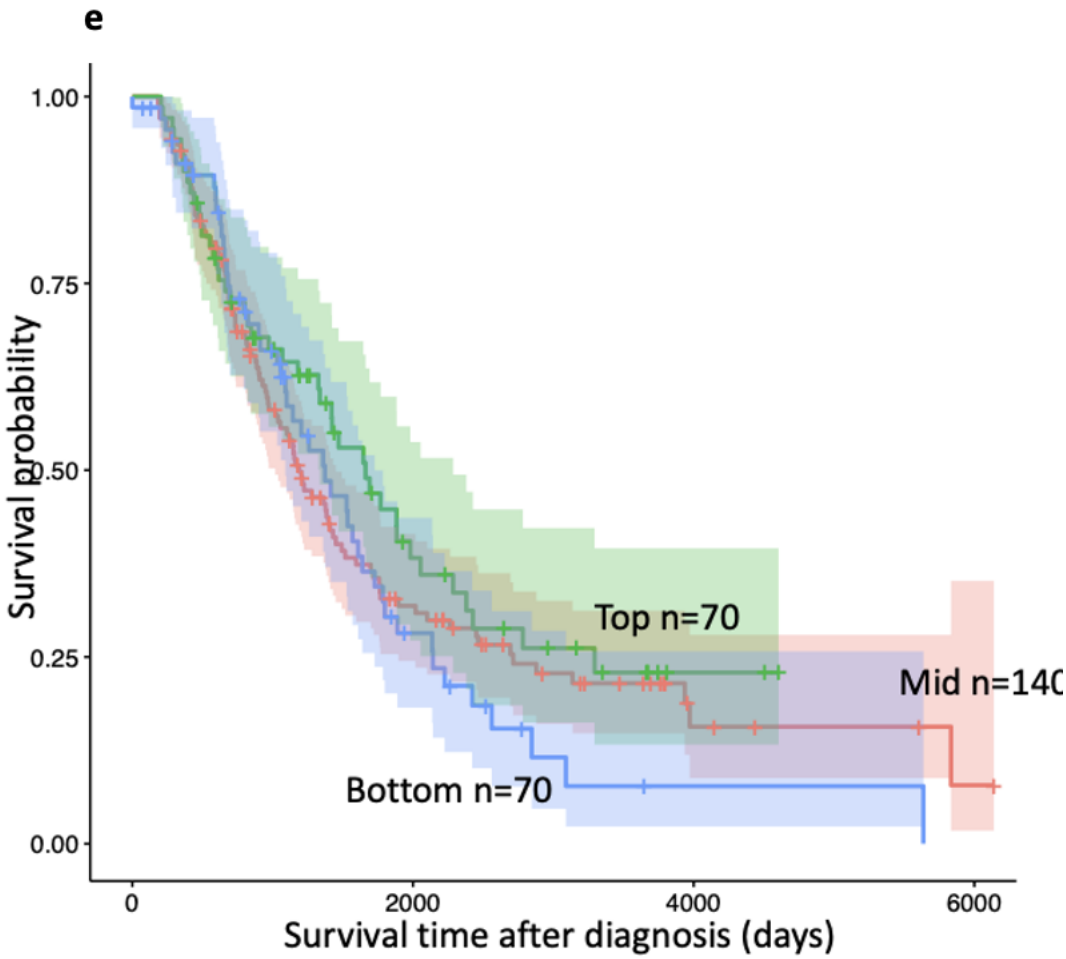

b
Survival probability
Survival time after diagnosis (days)
Top n=72
Mid n=137
Bottom n=71

c

Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.069	
Stage	N=280	1.46 (1.09-1.95)	0.011	
HRD	Absent N=118	Reference		
	Present N=162	0.48 (0.35-0.66)	0.001	
Cohort	SHGSOC N=81	Reference		
	AOCS N=80	1.66 (1.10-2.50)	0.015	
	BCCA N=59	0.82 (0.52-1.29)	0.39	
	TCGA N=31	1.53 (0.95-2.46)	0.084	
Severity of clusters	Middle N=137	Reference		
	Top N=72	0.90 (0.60-1.36)	0.611	
	Bottom N=71	1.11 (0.77-1.59)	0.583	

d




e
1.00
0.75
0.50
0.25
0.00
Survival probability
Top n=70
Mid n=140
Bottom n=70
0
2000
4000
6000
Survival time after diagnosis (days)

f

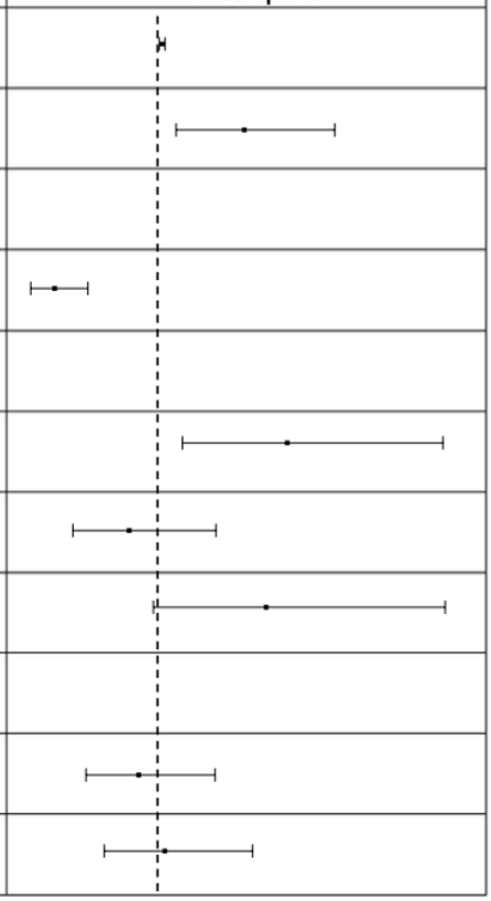
Condition		Hazard ratio	P value	Forest plot
Age	N=280	1.01 (1.00-1.03)	0.069	
Stage	N=280	1.46 (1.09-1.95)	0.011	
HRD	Absent N=118	Reference		
	Present N=162	0.48 (0.35-0.66)	0.001	
Cohort	SHGSOC N=81	Reference		
	AOCS N=80	1.66 (1.19-2.50)	0.015	
	BCCA N=59	0.82 (0.52-1.29)	0.59	
	TCGA N=31	1.52 (0.95-2.46)	0.084	
Cluster Severity	Mid N=137	Reference		
	Top N=72	0.90 (0.60-1.36)	0.611	
	Bottom N=71	1.11 (0.77-1.69)	0.583	

Figure 68 Severity of simple clustering and survival

a The distribution of severity of clusters measured by number of clusters in samples The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **b** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **a** the most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). **c** A Cox proportional hazards model comparing the survival of the samples most severely affected by clusters (top 25%) to the samples least effected by clusters (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort. **d** The distribution of severity of clusters measured by Mb covered by clusters in samples. The samples to the right of the red line are the top 25% most severity effected and the sample to the left Of the blue line are the 25% least severely impacted. **e** A Kaplan-Meier survival curve for time after diagnosis for three groups shown in **d** the

most severely impacted (green), the least severely impacted (blue), and remainder of the combined cohort (red). f A Cox proportional hazards model comparing the survival of the samples most severely affected by clusters (top 25%) to the samples least effected by clusters (bottom 25%) adjusting for age, stage at diagnosis, HRD status and sub-cohort.

Discussion

In the vast majority of studies involving cSV, only one type of cSV has been studied (Forment, Kaidi, and Jackson 2012; Z. Li et al. 2022), and some recent pan-cancer studies a few cSV types were studied (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisiation/Ins et al. 2020; Kim et al. 2020; Cortés-Ciriano, Lee, Xi, Jain, Jung, Yang, et al. 2020; Hadi et al. 2020). The work presented here provides the most in-depth analysis of a single uniformly procIssed cohort of a single cancer type, focusing on the impact of all currently reported cSV types and also the variation in severity of cSVs.

While the literature suggests that cSVs can have significant impacts on survival, this chapter demonstrates that the presence of any of the eight types of cSV studied did not lead to significantly worse or better survival in the combined cohort of HGSOC (Figure 51 – Figure 58) (Fontana et al. 2018; Forment, Kaidi, and Jackson 2012; Kloosterman, Koster, and Molenaar 2014; Magrangeas et al. 2011; Molenaar et al. 2012; Hadi et al. 2020; Qi, Li, and Sun 2016; Biswas et al. 2021). There are various factors that could account for these differences, including cancer type-specific effects, the use of different cSV definitions, and adjustments for different variables. Notably, HRD was found to be significantly mutually exclusive with cSV and has a protective impact on survival, so it is possible that genome-wide phenomena such as HRD could have confounded previous studies.

The effect of ecDNA on survival could not be separated from the cohort effect on survival as the majority of ecDNA found in samples were found in the AOCS cohort (Figure 58). The AOCS study, unlike the rest of the combined cohort, is enrchied for patients that did not respond to treatment and has been shown to have the worst survival of any of the sub-cohorts that make up the combined cohort (Figure 48). There was also no effect of cSV on progression free survival for any of the cSVs investigated (Figure 72 in Appendix).

The genes most frequently affected by cSV were olfactory and testis genes, which are often regarded as potential cancer biomarkers due to their low expression in most

tissues (Weber et al. 2018; Gjerstorff and Ditzel 2008). Genes whose expression is regulated by the PAX8 transcription factor were also repeatably found to have decreased expression in samples with cSV. This may suggest that treatments targeting PAX8 might be less effective in treating tumours with that cSV (C. Liu 2022).

Garsed et al. reported that an increased number of SV was associated with better survival time in a study of 126 HGSOC patients, including 60 patients with exceptionally long survival time after diagnosis (>10 years) (Garsed et al. 2022). In this chapter, the numbers of SV per sample demonstrated a similar trend but did not reach statistical significance (Figure 59).

Although the presence of cSV did not indicate better or worse survival, there was a trend that samples with more severe cSV had better survival. In cases where BFB severity was measured by size, this trend reached statistical significance. The analysis revealed that samples with the shortest BFB events had significantly worse survival compared to samples with longer BFB regions (**Figure 62**). The longest regions of chromothripsis also showed significantly better survival relative to other samples with chromothripsis. Similarly, samples with chromothripsis events containing fewer SV had significantly worse survival (**Figure 65**), again suggesting the most severe events of chromothripsis (containing more SV) are better for patient survival. When all cSV types were combined, samples with the most SV explained by cSV (in the top 25%) had significantly better survival than the remaining samples. Furthermore, the trend of larger complex regions being associated with better survival was observed when the total length of cSV per sample was tested. Overall this suggests that cSV burdens, whether measured by total genomic length or number of SVs involved, may be a valuable prognostic biomarker in HGSOC.

In chapter 4, it was noted that the majority of complexity in samples remains unexplained as most SV clusters (defined as unexpected clusters of simple SVs by Li et al (2020)) present in HGSOC samples were not explained by cSV. In this chapter it was

shown that the numbers of these clusters and the total length encompassed by these clusters did not have a significant impact on survival (Figure 68). It therefore seems that although these SV clusters have similar features in common with cSV, their variation in severity does not have the same impact on patient survival.

Chapter 6: Discussion

Complex structural variants (cSV) can result in massive genomic changes, potentially involving thousands of structural variants (SVs) across multi-megabase regions of the genome. Yet, the study of cSVs has usually been fragmented, often with a focus on a single or small number of cSV types. By investigating eight different cSV types simultaneously and utilizing the largest uniformly processed WGS cohort of HGSOC cases (n=324), this work has made several novel observations.

Pathways to Genomic Instability

There are two broadly divergent routes to the extreme structural variation seen in HGSOC, involving either the acquisition of HRD or the emergence of WGD with associated cSV (Figure 24). The results from chapter 3 demonstrate these divergent routes as non-mutually exclusive trends, where samples with HRD showed a depletion of cSV of all types, except for chromoplexy, (Table 6, Figure 27). Additionally, cSV types showed enrichment in samples with WGD, particularly chromothripsis and BFB (Table 5, Figure 26). WGD was also found to be significantly depleted in samples with HRD (Table 7). The depletion of WGD in samples with HRD, which is known to have improved survival, could explain why Bielski et al. reported worse survival for samples with WGD than samples without WGD pan-cancer (including 217 ovarian samples) (Bielski et al. 2018). Once adjusted for HRD and other covariates, WGD did not demonstrate a significant impact on survival in the combined cohort examined here (Figure 50).

Complex Structural Variants and Patient Survival

Once HRD and WGD had been adjusted for, the occurrence of cSV had no significant impact on survival in the combined cohort (Figure 51-Figure 58). In the literature, a positive or negative survival impact of the mere presence of cSV is often reported (Fontana et al. 2018; Forment, Kaidi, and Jackson 2012; Kloosterman, Koster, and Molenaar 2014; Magrangeas et al. 2011; Molenaar et al. 2012; Skuja et al. 2017;

Adelman and Martin 2021; Kim et al. 2020; Robert and Crasta 2022; Shah et al. 2021; Bailey et al. 2020). This discrepancy may be due to cancer type-specific events or not adjusting for the presence of genome-wide phenomena such as HRD, which is known to have a positive impact on survival in HGSOC.

In the specific case of ecDNA, any impact on survival could not be isolated from the effect of cohort membership (Figure 58). This is because, the majority of the samples with ecDNA in the combined cohort were identified in the AOCS sub-cohort which is biased to patients that did not respond to treatment (Figure 21). The AOCS sub-cohort had worse survival than the rest of the sub-cohorts (Figure 48) and in this work any survival impact of ecDNA was confounded by the cohort effect (Figure 58).

Although the presence of cSV by itself did not show any impact on survival, a difference was found between most and least severely impacted samples with cSVs (Figure 67). The severity of a cSV could be described by the number of SV involved or by the total genomic length of the cSV. –

Chromosome 19 - a Hotspot for cSVs in HGSOC

Chromosome 19 is a hotspot of cSV with chromothripsis, BFB and ecDNA all being enriched and having peaks near to the CCNE1 locus. While the enrichment of a specific cSV type on chromosomes has been previously reported, this is the first time that the enrichment of multiple cSV types on a single chromosome has been documented (Cortés-Ciriano, Lee, Xi, Jain, Jung, Zhang, et al. 2020).

BFB amplifications on chromosome 19 not only resulted in the amplification of the well-known oncogene CCNE1, but also led to a greater increase in CCNE1 copy number than simple duplications at the same locus (Figure 41). However, this did not translate to a significant impact on survival time for samples with BFB amplified CCNE1 (Figure 55).

Genomic Instability and the occurrence of cSVs

The genomic instability of a sample could be approximated by the number and size of SV and copy number changes (CNVs). Interestingly, it was observed that samples with a higher level of genomic instability did not necessarily have an increased likelihood of possessing a cSV once the SV accounted for by the cSV had been adjusted for. Therefore, samples with cSV are not just extreme outliers on the spectrum of structural complexity, but appear to represent distinct events, even on the background of frequent structural alterations seen in HGSOC.

Despite identifying a broad range of cSV, the majority of putative complex events (seen as significant SV clusters) in HGSOC remain unexplained. These unexplained SV clusters also exhibited key features of known cSV types, such as oscillation of copy number (Figure 32) which is unlikely to be generated by sequential rearrangement (Korbel and Campbell 2013). Overall, the number and size of these unexplained SV clusters did not have a significant impact on survival (Figure 68) but the large number of unexplained SV clusters suggests new cSV events may remain to be discovered.

The Future of Complex Structural Variant Research

The study of cSV is in its infancy. However, by looking at how research into better-understood genomic features has progressed, the future of cSV research can be anticipated. HRD is a well-understood molecular deficiency resulting in a well-defined pattern across the genome. Recent work has developed an accurate caller for HRD called HRDdetect, which specializes in calling HRD in breast cancer tissues (Davies et al. 2017), but an alternative algorithm known as CHORD claims to more accurately detect HRD across all tumour types (Nguyen et al. 2020). The detection methods for cSV are yet to be standardized in a single cancer type, let alone developed to encompass pan-cancer data. In addition, the algorithms used for identification of cSVs often rely upon heuristic approaches and arbitrary numerical thresholds, such as the threshold of 7 interleaved SV calls for the detection of chromothripsis used by the Shatterseek algorithm (Figure 6). Inevitably the thresholds that are appropriate in one tumour type may be less

appropriate in another.

HRD is not the only well understood source of genomic variation, and the COSMIC database has catalogued genome-wide signatures of SNV which have well understood underlying mechanisms, cancer specificities and even risk factors (Alexandrov et al. 2020; Otlu et al. 2022). Although cSV research is far from identifying signatures with this level of understanding, the recent development of the Starfish algorithm was an important step towards identifying mutational signatures for cSV research and linking underling mutational processes to cSVs (Bao et al. 2022).

Long Read Sequencing

The future of cSV research is intertwined with developments in DNA sequencing technologies such as long read sequencing (LRS). LRS technologies being developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) may advance cSV research in two ways: by allowing phasing of SV across the genome and by improving the mappability of reads and variant calling (van Belzen et al. 2021). LRS has been reported to be able to detect 3-4 times more germline SVs than short read sequencing (SRS) within normal human genomes (De Coster and Van Broeckhoven 2019; Audano et al. 2019; Chaisson et al. 2019). This increase is driven by an increased ability to map repetitive and GC rich regions of the genome, so is unlikely to translate to a 3-4 fold increase in the detection of somatic SV in cancer genomes (Chaisson et al. 2014; Audano et al. 2019; Chaisson et al. 2019), but the combination of LRS and SRS technologies will undoubtedly shed new light on cSV structures.

The work in this thesis focuses on utilising SRS to identify cSV. This technology utilises reads ~ 250bp long meaning that if two SV occur more than 250 bp apart, there is some degree of uncertainty as to whether they are affecting the same DNA molecule. Phasing of SV is now possible with LRS approaches due to the long length of LRS reads, with reports of reads up to 2.3Mb in length (Payne et al. 2019) and a report of 4.2MB by an internal ONT run which has not been peer reviewed (Nanopore 2021). These very long

reads are possible as the length of a run is only limited by the size of DNA molecules provided to the sequencing device (Amarasinghe et al. 2020). Phasing is important because cSV are essentially clusters of SV and CNV, and if the SV/CNV are on different DNA molecules then they are not clustered in reality (Lin et al. 2022). Phasing becomes more important and challenging as cancer genomes undergo WGD going from diploid to tetraploid, which we have seen is common in tumour types such as HGSOC.

LRS technologies also allow for the investigation of the relationship between cSV and methylation, which is currently poorly studied. However, LRS is not without is disadvantages, its base calling accuracy of ~85% is low compared to the ~99% accuracy of SRS. For this reason, approaches combining SRS and LRS have been utilized. For example, phased SV information has been used to identify a single "chromothripsis like event" in a sample from a cohort of 20 non-small cell lung cancer samples using a combination of SRS and LRS (Sakamoto et al. 2022). This study also reported that the chromothripsis like region was hypomethylated (Sakamoto et al. 2022) revealing a new impact of cSV on gene function.

Larger Cohorts

Larger cohorts of whole genome sequenced cancer samples will benefit all areas of cancer research including cSV research. As the size of WGS cancer cohorts continues to increase, it is likely that new cSV will become identifiable resolving more of the currently unexplained complexity. Additionally, it may be that cSV currently grouped into one type will be split into sub-groups as the large amount of heterogeneity may be masking patterns of complexity. By separating cSV from the heterogeneous groups currently defined into more uniform groupings, the survival impacts and functional effects of cSV may also become clearer and more interpretable.

Conclusion

This work has analysed the largest uniformly processed HGSOC cohort for a broad range of currently known cSV types. There were several novel results, including the discovery

of two broad routes to genomic variation, a hotspot of multiple cSV types on chromosome 19, and improved survival for samples with the most severe cSV events. This work has also highlighted that the majority of structural complexity within HGSOC remains unexplained.

As the phasing of SV becomes more accurate with LRS and cohorts of cancer samples increase in size, new cSVs are likely to be identified, and current cSV types may be subdivided into more consistent cSV types. However, these new classifications will not be useful unless they are consistently applied with standardized detection criteria appropriate for the cancer type being studied. As approaches to the detection of cSV become less heterogeneous between studies, their impact on cancer genomes may become easier to interpret.

References

100,000 Genomes Project Pilot Investigators, Damian Smedley, Katherine R. Smith, Antonio Martin, Ellen A. Thomas, Ellen M. McDonagh, Valentina Cipriani, et al. 2021. “100,000 –enomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report.” *The New England Journal of Medicine* 385 (20): 1868–80.

Acland, Mitchell, Georgia Arentz, Max Mussared, Fergus Whitehead, Peter Hoffmann, Manuela Klingler-Hoffmann, and Martin K. Oehler. 2020. “Proteomic Analysis of Pre-Invasive Serous Lesions of the Endometrium and Fallopian Tube Reveals Their Metastatic Potential.” *Frontiers in Oncology* 10 (December): 523989.

Adelman, Karen, and Benjamin J. E. Martin. 2021. “EcDNA Party Bus: Bringing the Enhancer to You.” *Molecular Cell* 81 (9): 1866–67.

Adler, Emily K., Rosario I. Corona, Janet M. Lee, Norma Rodriguez-Malave, Paulette Mhawech-Fauceglia, Heidi Sowter, Dennis J. Hazelett, Kate Lawrenson, and Simon A. Gayther. 2017. “The PAX8 Cistrome in Epithelial Ovarian Cancer.” *Oncotarget* 8 (65): 108316–32.

Ahmed, A. A., Etemadmoghadam, D., Temple, J., Lynch, A. G., Riad, M., Sharma, R., ... Brenton, J. D. (2010). Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *The Journal of Pathology*, 221(1), 49-56.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” *Nature* 578 (7793): 94–101.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30.

Ashby, Cody, Michael A. Bauer, Yan Wang, Christopher P. Wardell, Ruslana G. Tytarenko, Purvi Patel, Erin Flynt, et al. 2018. “Chromothripsis and Chromoplexy Are Associated with DNA Instability and Adverse Clinical Outcome in Multiple Myeloma.” *Blood* 132 (November): 408.

Ashby, Cody, Eileen M. Boyle, Brian A. Walker, Michael A. Bauer, Katie Rose Ryan, Judith Dent, Anjan Thakurta, Erin Flynt, Faith E. Davies, and Gareth Morgan. 2019. “Chromoplexy and Chromothripsis Are Important Prognostically in Myeloma and

Deregulate Gene Function By a Range of Mechanisms." *Blood* 134 (November): 3767.

Ashique, Sumel, Aakash Upadhyay, Ashish Garg, Neeraj Mishra, Afzal Hussain, Poonam Negi, Goh Bey Hing, et al. 2022. "Impact of EcDNA: A Mechanism That Directs Tumorigenesis in Cancer Drug Resistance-A Review." *Chemico-Biological Interactions* 363 (August): 110000.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." *Cell* 176 (3): 663-675.e19.

Baca, Sylvan C., Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, et al. 2013. "Punctuated Evolution of Prostate Cancer Genomes." *Cell* 153 (3): 666–77.

Bailey, C., M. J. Shoura, P. S. Mischel, and C. Swanton. 2020. "Extrachromosomal DNA—Relieving Heredity Constraints, Accelerating Tumour Evolution." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 31 (7): 884–93.

Bao, Lisui, Xiaoming Zhong, Yang Yang, and Lixing Yang. 2022. "Starfish Infers Signatures of Complex Genomic Rearrangements across Human Cancers." *Nature Cancer*, July, 1–13.

Barøy, Tale, Stine H. Kresse, Magne Skårn, Marianne Stabell, Russell Castro, Silje Lauvrak, Antonio Llombart-Bosch, Ola Myklebost, and Leonardo A. Meza-Zepeda. 2014. "Reexpression of LSAMP Inhibits Tumor Growth in a Preclinical Osteosarcoma Model." *Molecular Cancer* 13 (April): 93.

Bell D., A., M. Birrer Berchuck, D. W. Cramer J. Chien, F. Dao, R. Dhir, and Et Al. 2012. "Integrated Genomic Analyses of Ovarian Carcinoma," 0–7.

Belzen, Ianthe A. E. M. van, Alexander Schönhuth, Patrick Kemmeren, and Jayne Y. Hehir-Kwa. 2021. "Structural Variant Detection in Cancer Genomes: Computational Challenges and Perspectives for Precision Oncology." *Npj Precision Oncology* 5 (1): 1–11.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A

Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Berger, Michael F., Michael S. Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y. Sivachenko, Andrea Sboner, et al. 2011. "The Genomic Complexity of Primary Human Prostate Cancer." *Nature* 470 (7333): 214–20.

Bergsten, Tova M., Joanna E. Burdette, and Matthew Dean. 2020. "Fallopian Tube Initiation of High Grade Serous Ovarian Cancer and Ovarian Metastasis: Mechanisms and Therapeutic Implications." *Cancer Letters* 476 (April): 152–60.

Bianchi, Joy J., Valentine Murigneux, Marie Bedora-Faure, Chloé Lescale, and Ludovic Deriano. 2019. "Breakage-Fusion-Bridge Events Trigger Complex Genome Rearrangements and Amplifications in Developmentally Arrested T Cell Lymphomas." *Cell Reports* 27 (10): 2847-2858.e4.

Bielski, Craig M., Ahmet Zehir, Alexander V. Penson, Mark T. A. Donoghue, Walid Chatila, Joshua Armenia, Matthew T. Chang, et al. 2018. "Genome Doubling Shapes the Evolution and Prognosis of Advanced Cancers." *Nature Genetics* 50 (8): 1189–95.

Biswas, Subir, Gunjan Mandal, Kyle K. Payne, Carmen M. Anadon, Chandler D. Gatenbee, Ricardo A. Chaurio, Tara Lee Costich, et al. 2021. "IgA Transcytosis and Antigen Recognition Govern Ovarian Cancer Immunity." *Nature* 591 (7850): 464–70.

Boisselier, Blandine, Frédéric Dugay, Marc-Antoine Belaud-Rotureau, Anne Coutolleau, Emmanuel Garcion, Philippe Menei, Philippe Guardiola, and Audrey Rousseau. 2018. "Whole Genome Duplication Is an Early Event Leading to Aneuploidy in IDH-Wild Type Glioblastoma." *Oncotarget* 9 (89): 36017–28.

Brok, Wendie D. den, Kasmintan A. Schrader, Sophie Sun, Anna V. Tinker, Eric Yang Zhao, Samuel Aparicio, and Karen A. Gelmon. 2017. "Homologous Recombination Deficiency in Breast Cancer: A Clinical Review." *JCO Precision Oncology*, no. 1 (November): 1–13.

Cámara-Quílez, María, Aida Barreiro-Alonso, Ángel Vizoso-Vázquez, Esther Rodríguez-Belmonte, María Quindós-Varela, Mónica Lamas-Maceiras, and María Esperanza Cerdán. 2020. "The HMGB1-2 Ovarian Cancer Interactome. The Role of HMGB Proteins and Their Interacting Partners MIEN1 and NOP53 in Ovary Cancer and Drug-Response." *Cancers* 12 (9). https://doi.org/10.3390/cancers12092435.

Cameron, Daniel L., Jonathan Baber, Charles Shale, and Anthony T. Papenfuss. 2019. "GRIDSS , PURPLE , LINX : Unscrambling the Tumor Genome via Integrated

Analysis of Structural Variation and Copy Number Toolkit Description & Results Somatic Structural Variation (GRIDSS).”

Cameron, Daniel L., Jonathan Baber, Charles Shale, Anthony T. Papenfuss, Jose Espejo Valle-Inclan, Nicolle Besselink, Edwin Cuppen, and Peter Priestley. 2019. “GRIDSS, PURPLE, LINX: Unscrambling the Tumor Genome via Integrated Analysis of Structural Variation and Copy Number.” *BioRxiv*. https://doi.org/10.1101/781013.

Cameron, Daniel L., Leon Di Stefano, and Anthony T. Papenfuss. 2019. “Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software.” *Nature Communications* 10 (1): 1–11.

Cameron, Daniel L., Jan Schröder, Jocelyn Sietsma Penington, Hongdo Do, Ramyar Molania, Alexander Dobrovic, Terence P. Speed, and Anthony T. Papenfuss. 2017. “GRIDSS : Sensitive and Specific Genomic Rearrangement Detection Using Positional de Bruijn Graph Assembly,” 2050–60.

Carvalho, Claudia M. B., and James R. Lupski. 2016. “Mechanisms Underlying Structural Variant Formation in Genomic Disorders.” *Nature Reviews. Genetics* 17 (4): 224–38.

Carver, Brett S., Jennifer Tran, Anuradha Gopalan, Zhenbang Chen, Safa Shaikh, Arkaitz Carracedo, Andrea Alimonti, et al. 2009. “Aberrant ERG Expression Cooperates with Loss of PTEN to Promote Cancer Progression in the Prostate.” *Nature Genetics* 41 (5): 619–24.

Cen, Yixuan, Yifeng Fang, Yan Ren, Shiyuan Hong, Weiguo Lu, and Junfen Xu. 2022. “Global Characterization of Extrachromosomal Circular DNAs in Advanced High Grade Serous Ovarian Cancer.” *Cell Death & Disease* 13 (4): 1–10.

Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2014. “Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing.” *Nature* 517 (7536): 608–11.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. “Multi-Platform Discovery of

Haplotype-Resolved Structural Variation in Human Genomes." *Nature Communications* 10 (1): 1–16.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. "Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications." *Bioinformatics* 32 (8): 1220–22.

Cheung, Hiu Wing, Glenn S. Cowley, Barbara A. Weir, Jesse S. Boehm, Scott Rusin, Justine A. Scott, Alexandra East, et al. 2011. "Systematic Investigation of Genetic Vulnerabilities across Cancer Cell Lines Reveals Lineage-Specific Dependencies in Ovarian Cancer." *Proceedings of the National Academy of Sciences of the United States of America* 108 (30): 12372–77.

Chiang, Colby, Jessie C. Jacobsen, Carl Ernst, Carrie Hanscom, Ian Blumenthal, Ryan E. Mills, Andrew Kirby, et al. 2012. "Complex Reorganization and Predominant Non-Homologous Repair Following Chromosomal Breakage in Karyotypically Balanced Germline Rearrangements and Transgenic Integration" 44 (4): 1–18.

Chiara, Maria, Fontana Giovanni, Marconi Jelena, D. Milosevic Feenstra, Eugenio Fonzi, Cristina Papayannidis, Andrea Ghelli, et al. 2018. "Chromothripsis in Acute Myeloid Leukemia : Biological Features and Impact on Survival." *Leukemia*, 1609–20.

Church, G. M., and W. Gilbert. 1984. "Genomic Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 81 (7): 1991–95.

Cortés-Ciriano, Isidro, Jake June Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. "Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing." *Nature Genetics* 52 (3): 331–41.

Cortés-Ciriano, Isidro, June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. "Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing," LB-378-LB-378.

Cosenza, Marco Raffaele, Bernardo Rodriguez-Martin, and Jan O. Korbel. 2022. "Structural Variation in Cancer: Role, Prevalence, and Mechanisms." Annual

Review of Genomics and Human Genetics, June. https://doi.org/10.1146/annurev-genom-120121-101149.

Coutelier, Marie, Manuel Holtgrewe, Marten Jäger, Ricarda Flöttman, Martin A. Mensah, Malte Spielmann, Peter Kfrawitz, Denise Horn, Dieter Beule, and Stefan Mundlos. 2022. "Combining Callers Improves the Detection of Copy Number Variants from Whole-Genome Sequencing." *European Journal of Human Genetics: EJHG* 30 (2): 178–86.

Cox, D., C. Yuncken, and A. I. Spriggs. 1965. "MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD." *The Lancet* 1 (7402): 55–58.

Crum, Christopher P. 2009. "Intercepting Pelvic Cancer in the Distal Fallopian Tube: Theories and Realities." *Molecular Oncology* 3 (2): 165–70.

Cunha Colombo Bonadio, Renata Rodrigues da, Rodrigo Nogueira Fogace, Vanessa Costa Miranda, and Maria Del Pilar Estevez Diz. 2018. "Homologous Recombination Deficiency in Ovarian Cancer: A Review of Its Epidemiology and Management." *Clinics* 73 (suppl 1): e450s.

Davies, Helen, Dominik Glodzik, Sandro Morganella, Lucy R. Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, et al. 2017. "HRDetect Is a Predictor of BRCA1 and BRCA2 Deficiency Based on Mutational Signatures." *Nature Medicine* 23 (4): 517–25.

De Coster, Wouter, and Christine Van Broeckhoven. 2019. "Newest Methods for Detecting Structural Variations." *Trends in Biotechnology* 37 (9): 973–82.

Deshpande, Viraj, Jens Luebeck, Nam Phuong D. Nguyen, Mehrdad Bakhtiari, Kristen M. Turner, Richard Schwab, Hannah Carter, Paul S. Mischel, and Vineet Bafna. 2019. "Exploring the Landscape of Focal Amplifications in Cancer Using AmpliconArchitect." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-018-08200-y.

Di Palma, Tina, Valeria Lucci, Tiziana de Cristofaro, Maria Grazia Filippone, and Mariastella Zannini. 2014. "A Role for PAX8 in the Tumorigenic Phenotype of Ovarian Cancer Cells." *BMC Cancer* 14 (April): 292.

Difilippantonio, Michael J., Simone Petersen, Hua Tang Chen, Roger Johnson, Maria Jasin, Roland Kanaar, Thomas Ried, and André Nussenzweig. 2002. "Evidence for

Replicative Repair of DNA Double-Strand Breaks Leading to Oncogenic Translocation and Gene Amplification." *The Journal of Experimental Medicine* 196 (4): 469–80.

Drews, Ruben M., Barbara Hernando, Maxime Tarabichi, Kerstin Haase, Tom Lesluyes, Philip S. Smith, Lena Morrill Gavarró, et al. 2022. "A Pan-Cancer Compendium of Chromosomal Instability." *Nature* 606 (7916): 976–83.

Durmaz, Asude Alpman, Emin Karaca, Urszula Demkow, Gokce Toruner, Jacqueline Schoumans, and Ozgur Cogulu. 2015. "Evolution of Genetic Techniques: Past, Present, and Beyond." *BioMed Research International* 2015 (March): 461524.

Elias, Kevin M., Megan M. Emori, Thomas Westerling, Henry Long, Anna Budina-Kolomets, Fugen Li, Emily MacDuffie, et al. 2016. "Epigenetic Remodeling Regulates Transcriptional Changes between Ovarian Cancer and Benign Precursors." *JCI Insight* 1 (13). https://doi.org/10.1172/jci.insight.87988.

Engen-van Grunsven, Adriana C. H. van, Marjolein P. Baar, Rolph Pfundt, Jos Rijntjes, Heidi V. N. Küsters-Vandevelde, Ann-Laure Delbecq, Jan E. Keunen, et al. 2015. "Whole-Genome Copy-Number Analysis Identifies New Leads for Chromosomal Aberrations Involved in the Oncogenesis and Metastatic Behavior of Uveal Melanomas." *Melanoma Research* 25 (3): 200–209.

Ewing, Ailith, Alison Meynert, Michael Churchman, Graeme R. Grimes, Robert L. Hollis, C. Simon Herrington, Tzyvia Rye, et al. 2020. "Structural Variants at the BRCA1/2 Loci Are a Common Source of Homologous Repair Deficiency in High Grade Serous Ovarian Carcinoma." *BioRxiv*. https://doi.org/10.1101/2020.05.11.088278.

Farinella, Federica, Mario Merone, Luca Bacco, Adriano Capirchio, Massimo Ciccozzi, and Daniele Caligiore. 2022. "Machine Learning Analysis of High-Grade Serous Ovarian Cancer Proteomic Dataset Reveals Novel Candidate Biomarkers." *Scientific Reports* 12 (1): 3041.

Fontana, Maria Chiara, Giovanni Marconi, Jelena D. Milosevic Feenstra, Eugenio Fonzi, Cristina Papayannidis, Andrea Ghelli Luserna di Rorá, Antonella Padella, et al. 2018. "Chromothripsis in Acute Myeloid Leukemia: Biological Features and Impact on Survival." *Leukemia* 32 (7): 1609–20.

Forment, Josep V., Abderrahmane Kaidi, and Stephen P. Jackson. 2012. “Chromothripsis and Cancer: Causes and Consequences of Chromosome Shattering.” *Nature Reviews. Cancer* 12 (10): 663–70.

Fratta, Elisabetta, Sandra Coral, Alessia Covre, Giulia Parisi, Francesca Colizzi, Riccardo Danielli, Hugues Jean Marie Nicolay, Luca Sigalotti, and Michele Maio. 2011. “The Biology of Cancer Testis Antigens: Putative Function, Regulation and Therapeutic Potential.” *Molecular Oncology* 5 (2): 164–82.

Frey, Melissa K., and Bhavana Pothuri. 2017. “Homologous Recombination Deficiency (HRD) Testing in Ovarian Cancer Clinical Practice: A Review of the Literature.” *Gynecologic Oncology Research and Practice* 4 (February): 4.

Fujiwara, Takeshi, Madhavi Bandi, Masayuki Nitta, Elena V. Ivanova, Roderick T. Bronson, and David Pellman. 2005. “Cytokinesis Failure Generating Tetraploids Promotes Tumorigenesis in P53-Null Cells.” *Nature* 437 (7061): 1043–47.

Gabrielaite, Migle, Mathias Husted Torp, Malthe Sebro Rasmussen, Sergio Andreu-Sánchez, Filipe Garrett Vieira, Christina Bligaard Pedersen, Savvas Kinalis, et al. 2021. “A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data.” *Cancers* 13 (24). https://doi.org/10.3390/cancers13246283.

Garsed, Dale W., Owen J. Marshall, Vincent D. A. Corbin, Arthur Hsu, Leon Di Stefano, Jan Schröder, Jason Li, et al. 2014. “The Architecture and Evolution of Cancer Neochromosomes.” *Cancer Cell* 26 (5): 653–67.

Garsed, Dale W., Ahwan Pandey, Sian Fereday, Catherine J. Kennedy, Kazuaki Takahashi, Kathryn Alsop, Phineas T. Hamilton, et al. 2022. “The Genomic and Immune Landscape of Long-Term Survivors of High-Grade Serous Ovarian Cancer.” *Nature Genetics* 54 (12): 1853–64.

Gemble, Simon, René Wardenaar, Kristina Keuper, Nishit Srivastava, Maddalena Nano, Anne-Sophie Macé, Andréa E. Tijhuis, et al. 2022. “Genetic Instability from a Single S Phase after Whole-Genome Duplication.” *Nature* 604 (7904): 146–51.

Gisselsson, D., L. Pettersson, M. Höglund, M. Heidenblad, L. Gorunova, J. Wiegant, F. Mertens, P. Dal Cin, F. Mitelman, and N. Mandahl. 2000. “Chromosomal Breakage-Fusion-Bridge Events Cause Genetic Intratumor Heterogeneity.”

Proceedings of the National Academy of Sciences of the United States of America 97 (10): 5357–62.

Gjerstorff, M. F., and H. J. Ditzel. 2008. "An Overview of the GAGE Cancer/Testis Antigen Family with the Inclusion of Newly Identified Members." *Tissue Antigens* 71 (3): 187–92.

Glassman, A. B. 2000. "Chromosomal Abnormalities in Acute Leukemias." *Clinics in Laboratory Medicine* 20 (1): 39–48.

Goh, H. S., J. Yao, and D. R. Smith. 1995. "P53 Point Mutation and Survival in Colorectal Cancer Patients." *Cancer Research* 55 (22): 5217–21.

Gorski, Justin W., Frederick R. Ueland, and Jill M. Kolesar. 2020. "CCNE1 Amplification as a Predictive Biomarker of Chemotherapy Resistance in Epithelial Ovarian Cancer." *Diagnostics (Basel, Switzerland)* 10 (5). https://doi.org/10.3390/diagnostics10050279.

Govind, Shaylan K., Amin Zia, Pablo H. Hennings-Yeomans, John D. Watson, Michael Fraser, Catalina Anghel, Alexander W. Wyatt, et al. 2014. "ShatterProof: Operational Detection and Quantification of Chromothripsis." *BMC Bioinformatics* 15 (1). https://doi.org/10.1186/1471-2105-15-78.

Gresham, David, Maitreya J. Dunham, and David Botstein. 2008. "Comparing Whole Genomes Using DNA Microarrays." *Nature Reviews. Genetics* 9 (4): 291–302.

Hadi, Kevin, Xiaotong Yao, Julie M. Behr, Aditya Deshpande, Charalampos Xanthopoulakis, Huasong Tian, Sarah Kudman, et al. 2020. "Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs." *Cell* 183 (1): 197-210.e32.

Hanahan, Douglas, and Robert A. Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100 (1): 57–70.

Hart, Jonathan R., Yaoyang Zhang, Lujian Liao, Lynn Ueno, Lisa Du, Marloes Jonkers, John R. Yates, and Peter K. Vogt. 2015. "The Butterfly Effect in Cancer: A Single Base Mutation Can Remodel the Cell." *Proceedings of the National Academy of Sciences* 112 (4): 1131–36.

He, Quanze, Quanyuan He, Xiaohui Liu, Youheng Wei, Suqin Shen, Xiaohui Hu, Qiao Li, Xiangwen Peng, Lin Wang, and Long Yu. 2014. "Genome-Wide Prediction of

Cancer Driver Genes Based on SNP and Cancer SNV Data." *American Journal of Cancer Research* 4 (4): 394–410.

Hoheisel, Jörg D. 2006. "Microarray Technology: Beyond Transcript Profiling and Genotype Analysis." *Nature Reviews. Genetics* 7 (3): 200–210.

Hollis, Robert L., Alison M. Meynert, Caroline O. Michie, Tzyvia Rye, Michael Churchman, Amelia Hallas-Potts, Ian Croy, et al. 2022. "Multiomic Characterization of High-Grade Serous Ovarian Carcinoma Enables High-Resolution Patient Stratification." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 28 (16): 3546–56.

Hoppe, Michal M., Raghav Sundar, David S. P. Tan, and Anand D. Jeyasekharan. 2018. "Biomarkers for Homologous Recombination Deficiency in Cancer." *Journal of the National Cancer Institute* 110 (7): 704–13.

Hung, King L., Kathryn E. Yost, Liangqi Xie, Quanming Shi, Konstantin Helmsauer, Jens Luebeck, Robert Schöpflin, et al. 2021. "EcDNA Hubs Drive Cooperative Intermolecular Oncogene Expression." *Nature* 600 (7890): 731–36.

Ilić, Mila, Irene C. Zaalberg, Jonne A. Raaijmakers, and René H. Medema. 2022. "Life of Double Minutes: Generation, Maintenance, and Elimination." *Chromosoma* 131 (3): 107–25.

Imielinski, Marcin, Guangwu Guo, and Matthew Meyerson. 2017. "Insertions and Deletions Target Lineage-Defining Genes in Human Cancers." *Cell* 168 (3): 460-472.e14.

Jacobs, Ian J., Usha Menon, Andy Ryan, Aleksandra Gentry-Maharaj, Matthew Burnell, Jatinderpal K. Kalsi, Nazar N. Amso, et al. 2016. "Ovarian Cancer Screening and Mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A Randomised Controlled Trial." *The Lancet* 387 (10022): 945–56.

Jay, Ash, Diedre Reitz, Satoshi H. Namekawa, and Wolf-Dietrich Heyer. 2021. "Cancer Testis Antigens and Genomic Instability: More than Immunology." *DNA Repair* 108 (December): 103214.

Jemal, Ahmedin, Rebecca Siegel, Jiaquan Xu, and Elizabeth Ward. 2010. "Cancer Statistics, 2010." *CA: A Cancer Journal for Clinicians* 60 (5): 277–300.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Kanehisa, Minoru. 2019. “Toward Understanding the Origin and Evolution of Cellular Organisms.” *Protein Science: A Publication of the Protein Society* 28 (11): 1947–51.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. 2023. “KEGG for Taxonomy-Based Analysis of Pathways and Genomes.” *Nucleic Acids Research* 51 (D1): D587–92.

Karst, Alison M., and Ronny Drapkin. 2010. “Ovarian Cancer Pathogenesis: A Model in Evolution.” *Journal of Oncology* 2010: 932371.

Kim, Hoon, Nam Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, et al. 2020. “Extrachromosomal DNA Is Associated with Oncogene Amplification and Poor Outcome across Multiple Cancers.” *Nature Genetics* 52 (9): 891–97.

King, Jennifer C., Jin Xu, John Wongvipat, Haley Hieronymus, Brett S. Carver, David H. Leung, Barry S. Taylor, et al. 2009. “Cooperativity of TMPRSS2-ERG with PI3-Kinase Pathway Activation in Prostate Oncogenesis.” *Nature Genetics*.

Kinsella, Marcus, and Vineet Bafna. 2012. “Combinatorics of the Breakage-Fusion-Bridge Mechanism.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (6): 662–78.

Kinsella, Marcus, Anand Patel, and Vineet Bafna. 2014. “The Elusive Evidence for Chromothripsis.” *Nucleic Acids Research* 42 (13): 8231–42.

Klaasen, Sjoerd J., My Anh Truong, Richard H. van Jaarsveld, Isabella Koprivec, Valentina Štimac, Sippe G. de Vries, Patrik Risteski, et al. 2022. “Nuclear Chromosome Locations Dictate Segregation Error Frequencies.” *Nature* 607 (7919): 604–9.

Kloosterman, Wigard P., Jan Koster, and Jan J. Molenaar. 2014. “Prevalence and Clinical Implications of Chromothripsis in Cancer Genomes.” *Current Opinion in Oncology* 26 (1): 64–72.

Kohl, N. E., N. Kanda, R. R. Schreck, G. Bruns, S. A. Latt, F. Gilbert, and F. W. Alt. 1983. “Transposition and Amplification of Oncogene-Related Sequences in Human Neuroblastomas.” *Cell* 35 (2 Pt 1): 359–67.

Konstantinopoulos, Panagiotis A., Raphael Ceccaldi, Geoffrey I. Shapiro, and Alan D. D'Andrea. 2015. "Homologous Recombination Deficiency: Exploiting the Fundamental Vulnerability of Ovarian Cancer." *Cancer Discovery* 5 (11): 1137–54.

Korbel, Jan O., and Peter J. Campbell. 2013. "Criteria for Inference of Chromothripsis in Cancer Genomes." *Cell* 152 (6): 1226–36.

Kovtun, Irina V., Stephen J. Murphy, Sarah H. Johnson, John C. Cheville, and George Vasmatzis. 2015. "Chromosomal Catastrophe Is a Frequent Event in Clinically Insignificant Prostate Cancer." *Oncotarget* 6 (30): 29087–96.

Kresse, Stine H., Hege O. Ohnstad, Erik B. Paulsen, Bodil Bjerkehagen, Karoly Szuhai, Massimo Serra, Karl-Ludwig Schaefer, Ola Myklebost, and Leonardo A. Meza-Zepeda. 2009. "LSAMP, a Novel Candidate Tumor Suppressor Gene in Human Osteosarcomas, Identified by Array Comparative Genomic Hybridization." *Genes, Chromosomes & Cancer* 48 (8): 679–93.

Kroeger, Paul T., Jr, and Ronny Drapkin. 2017. "Pathogenesis and Heterogeneity of Ovarian Cancer." *Current Opinion in Obstetrics & Gynecology* 29 (1): 26–34.

Kumar, Rajesh, Gandharva Nagpal, Vinod Kumar, Salman Sadullah Usmani, Piyush Agrawal, and Gajendra P. S. Raghava. 2019. "HumCFS: A Database of Fragile Sites in Human Chromosomes." *BMC Genomics* 19 (Suppl 9): 985.

Labidi-Galy, S. Intidhar, Eniko Papp, Dorothy Hallberg, Noushin Niknafs, Vilmos Adleff, Michael Noe, Rohit Bhattacharya, et al. 2017. "High Grade Serous Ovarian Carcinomas Originate in the Fallopian Tube." *Nature Communications* 8 (1): 1–11.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology* 9 (8): e1003118.

Lens, Susanne M. A., and René H. Medema. 2018. "Cytokinesis Defects and Cancer." *Nature Reviews. Cancer* 19 (1): 32–45.

Levine, Michelle S., and Andrew J. Holland. 2018. “The Impact of Mitotic Errors on Cell Proliferation and Tumorigenesis.” *Genes & Development* 32 (9–10): 620–38.

Li, Xuan, and Wolf-Dietrich Heyer. 2008. “Homologous Recombination in DNA Repair and DNA Damage Tolerance.” *Cell Research* 18 (1): 99–113.

Li, Yilong, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, et al. 2020. “Patterns of Somatic Structural Variation in Human Cancer Genomes.” *Nature* 578 (7793): 112–21.

Li, Zesheng, Bo Wang, Hao Liang, and Lei Han. 2022. “Pioneering Insights of Extrachromosomal DNA (EcDNA) Generation, Action and Its Implications for Cancer Therapy.” *International Journal of Biological Sciences* 18 (10): 4006–25.

Lin, Jyun-Hong, Liang-Chi Chen, Shu-Chi Yu, and Yao-Ting Huang. 2022. “LongPhase: An Ultra-Fast Chromosome-Scale Phasing Algorithm for Small and Large Variants.” *Bioinformatics* 38 (7): 1816–22.

Lisio, Michael Antony, Lili Fu, Alicia Goyeneche, Zu Hua Gao, and Carlos Telleria. 2019. “High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints.” *International Journal of Molecular Sciences* 20 (4). https://doi.org/10.3390/ijms20040952.

Liu, Caihong. 2022. “Curcumol Targeting PAX8 Inhibits Ovarian Cancer Cell Migration and Invasion and Increases Chemotherapy Sensitivity of Niraparib.” *Journal of Oncology* 2022 (May): 3941630.

Liu, Guo, Joshua Stevens, Steven Horne, Batoul Abdallah, Karen Ye, Steven Bremer, Christine Ye, David J. Chen, and Henry Heng. 2014. “Genome Chaos: Survival Strategy during Crisis.” *Cell Cycle* 13 (4): 528–37.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2,” 1–21.

Magrangeas, Florence, Hervé Avet-Loiseau, Nikhil C. Munshi, and Stéphane Minvielle. 2011. “Chromothripsis Identifies a Rare and Aggressive Entity among Newly Diagnosed Multiple Myeloma Patients.” *Blood* 118 (3): 675–78.

Maher, A., and Richard K. Wilson. 2013. “Chromothripsis and Human Disease: Piecing Together the Shattering Process.” *Cell* 148 (0): 29–32.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. “Structural Variant Calling: The Long and the Short of It.” *Genome Biology* 20 (1): 246.

McClintock, B. 1941. “The Stability of Broken Ends of Chromosomes in Zea Mays.” *Genetics* 26 (2): 234–82.

Mei, Jie, Huixiang Tian, Hsuan-Shun Huang, Nayiyuan Wu, Yu-Ligh Liou, Tang-Yuan Chu, Jing Wang, and Wei Zhang. 2023. “CCNE1 Is a Potential Target of Metformin for Tumor Suppression of Ovarian High-Grade Serous Carcinoma.” *Cell Cycle* 22 (1): 85–99.

Menon, Usha, Andy Ryan, Jatinderpal Kalsi, Aleksandra Gentry-Maharaj, Anne Dawnay, Mariam Habib, Sophia Apostolidou, et al. 2015. “Risk Algorithm Using Serial Biomarker Measurements Doubles the Number of Screen-Detected Cancers Compared with a Single-Threshold Rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 33 (18): 2062–71.

Menon, U., Gentry-Maharaj, A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., et al. (2021). Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *The Lancet*, 397(10290), 2182-2193

Mermel, Craig H., Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. 2011. “GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers.” *Genome Biology* 12 (4): R41.

Metzker, Michael L. 2009. “Sequencing Technologies — the next Generation.” *Nature Reviews. Genetics* 11 (1): 31–46.

Mina, Marco, Arvind Iyer, Daniele Tavernari, Franck Raynaud, and Giovanni Ciriello. 2020. “Discovering Functional Evolutionary Dependencies in Human Cancers.” *Nature Genetics* 52 (11): 1198–1207.

Mina, Marco, Franck Raynaud, Daniele Tavernari, Elena Battistello, Stephanie Sungalee,

Sadegh Saghafinia, Titouan Laessle, et al. 2017. “Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies.” *Cancer Cell* 32 (2): 155-168.e6.

Mitchell, Thomas J., Samra Turajlic, Andrew Rowan, David Nicol, James H. R. Farmery, Tim O’Brien, Inigo Martincorena, et al. 2018. “Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal.” *Cell* 173 (3): 611-623.e17.

Mittelman, David, and John H. Wilson. 2013. “The Fractured Genome of HeLa Cells.” *Genome Biology*.

Molenaar, Jan J., Jan Koster, Danny A. Zwijnenburg, Peter Van Sluis, Linda J. Valentijn, Ida Van Der Ploeg, Mohamed Hamdi, et al. 2012. “Sequencing of Neuroblastoma Identifies Chromothripsis and Defects in Neuritogenesis Genes.” *Nature* 483 (7391): 589–93.

Moorhead, P. S., P. C. Nowell, W. J. Mellman, D. M. Battips, and D. A. Hungerford. 1960. “Chromosome Preparations of Leukocytes Cultured from Human Peripheral Blood.” *Experimental Cell Research* 20 (September): 613–16.

Morden, Claire R., Ally C. Farrell, Mirka Sliwowski, Zelda Lichtensztejn, Alon D. Altman, Mark W. Nachtigal, and Kirk J. McManus. 2021. “Chromosome Instability Is Prevalent and Dynamic in High-Grade Serous Ovarian Cancer Patient Samples.” *Gynecologic Oncology* 161 (3): 769–78.

Moreno-Cabrera, José Marcos, Jesús Del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro, and Bernat Gel. 2020. “Evaluation of CNV Detection Tools for NGS Panel Data in Genetic Diagnostics.” *European Journal of Human Genetics: EJHG* 28 (12): 1645–55.

Nanopore, Oxford. 2021. “In Combination with @circulomics #Nanobind Kits, the New @nanopore Ultra-Long DNA Sequencing Kit Enables Read Lengths of up to 4.2 Mb and Maximises the Quantity of Ultra-Long Fragments — Now Available in Store. Read More: Https://T.Co/NMpSTxMraS Pic.twitter.com/0tpreiqh2c.” Twitter. March 25, 2021. https://twitter.com/nanopore/status/1374997355154575366?lang=en.

Nath, Aritro, Patrick A. Cosgrove, Hoda Mirsafian, Elizabeth L. Christie, Lance Pflieger, Benjamin Copeland, Sumana Majumdar, et al. 2021. “Evolution of Core

Archetypal Phenotypes in Progressive High Grade Serous Ovarian Cancer." *Nature Communications* 12 (1): 1–16.

Nguyen, Luan, John W. M. Martens, Arne Van Hoeck, and Edwin Cuppen. 2020. "Pan-Cancer Landscape of Homologous Recombination Deficiency." *Nature Communications* 11 (1): 1–12.

Nikolaev, Sergey, Federico Santoni, Marco Garieri, Periklis Makrythanasis, Emilie Falconnet, Michel Guipponi, Anne Vannier, et al. 2014. "Extrachromosomal Driver Mutations in Glioblastoma and Low-Grade Glioma." *Nature Communications* 5 (1): 1–7.

Noguchi, Shuhei, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, et al. 2017. "FANTOM5 CAGE Profiles of Human and Mouse Samples." *Scientific Data* 4 (August): 170112.

Norris, Eric J., Wendell D. Jones, Marius D. Surleac, Andrei J. Petrescu, Darla Destephanis, Qing Zhang, Issam Hamadeh, et al. 2018. "Clonal Lineage of High Grade Serous Ovarian Cancer in a Patient with Neurofibromatosis Type 1." *Gynecologic Oncology Reports* 23 (February): 41–44.

Northcott, Paul A., David J. H. Shih, John Peacock, Livia Garzia, A. Sorana Morrissy, Thomas Zichner, Adrian M. Stútz, et al. 2012. "Subgroup-Specific Structural Variation across 1,000 Medulloblastoma Genomes." *Nature* 487 (7409): 49–56.

Notta, Faiyaz, Michelle Chan-Seng-Yue, Mathieu Lemire, Yilong Li, Gavin W. Wilson, Ashton A. Connor, Robert E. Denroche, et al. 2016. "A Renewed Model of Pancreatic Cancer Evolution Based on Genomic Rearrangement Patterns." *Nature* 538 (7625): 378–82.

Nowell, P. C., and D. A. Hungerford. 1960. "Chromosome Studies on Normal and Leukemic Human Leukocytes." *Journal of the National Cancer Institute* 25 (July): 85–109.

Otlu, Burcak, Marcos Díaz-Gay, Ian Vermes, Erik N. Bergstrom, Mark Barnes, and Ludmil B. Alexandrov. 2022. "Topography of Mutational Signatures in Human Cancer." *BioRxiv*. https://doi.org/10.1101/2022.05.29.493921.

Ozaki, K., T. Enomoto, K. Yoshino, M. Fujita, G. S. Buzard, K. Kawano, M. Yamasaki, and Y. Murata. 2001. “Impaired FHIT Expression Characterizes Serous Ovarian Carcinoma.” *British Journal of Cancer* 85 (2): 247–54.

Patch, Ann-Marie, Elizabeth L. Christie, Dariush Etemadmoghadam, Dale W. Garsed, Joshy George, Sian Fereday, Katia Nones, et al. 2015. “Whole-Genome Characterization of Chemoresistant Ovarian Cancer.” https://doi.org/10.1038/nature14410.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. 2019. “BulkVis: A Graphical Viewer for Oxford Nanopore Bulk FAST5 Files.” *Bioinformatics* 35 (13): 2193–98.

Pennington, Kathryn P., Tom Walsh, Maria I. Harrell, Ming K. Lee, Christopher C. Pennil, Mara H. Rendi, Anne Thornton, et al. 2014. “Germline and Somatic Mutations in Homologous Recombination Genes Predict Platinum Response and Survival in Ovarian, Fallopian Tube, and Peritoneal Carcinomas.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 20 (3): 764–75.

Petersen, Shariska, Andrew J. Wilson, Jeff Hirst, Katherine F. Roby, Oluwole Fadare, Marta A. Crispens, Alicia Beeghly-Fadiel, and Dineo Khabele. 2020. “CCNE1 and BRD4 Co-Amplification in High-Grade Serous Ovarian Cancer Is Associated with Poor Clinical Outcomes.” *Gynecologic Oncology* 157 (2): 405–10.

Przybycin, Christopher G., Robert J. Kurman, Brigitte M. Ronnett, Ie-Ming Shih, and Russell Vang. 2010. “Are All Pelvic (Nonuterine) Serous Carcinomas of Tubal Origin?” *The American Journal of Surgical Pathology* 34 (10): 1407–16.

Purshouse, Karin, Elias T. Friman, Shelagh Boyle, Pooran Singh Dewari, Vivien Grant, Alhafidz Hamdan, Gillian M. Morrison, et al. 2022. “Oncogene Expression from Extrachromosomal DNA Is Driven by Copy Number Amplification and Does Not Require Spatial Clustering.” *BioRxiv*. https://doi.org/10.1101/2022.01.29.478046.

Qi, Xuanchang, Xuechang Li, and Xiuxia Sun. 2016. “Reduced Expression of Polymeric Immunoglobulin Receptor (PIgR) in Nasopharyngeal Carcinoma and Its Correlation with Prognosis.” *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine* 37 (8): 11099–104.

Quinton, Ryan J., Amanda DiDomizio, Marc A. Vittoria, Kristýna Kotýnková, Carlos J. Ticas, Sheena Patel, Yusuke Koga, et al. 2021. “Whole-Genome Doubling Confers Unique Genetic Vulnerabilities on Tumour Cells.” *Nature* 590 (7846): 492–97.

Raab, Monika, Nene F. Kobayashi, Sven Becker, Elisabeth Kurunci-Csacsko, Andrea Krämer, Klaus Strebhardt, and Mourad Sanhaji. 2020. “Boosting the Apoptotic Response of High-Grade Serous Ovarian Cancers with CCNE1 Amplification to Paclitaxel in Vitro by Targeting APC/C and the pro-Survival Protein MCL-1.” *International Journal of Cancer. Journal International Du Cancer* 146 (4): 1086–98.

Rausch, Tobias, David T. W. Jones, Marc Zapatka, Adrian M. Stütz, Joachim Weischenfeldt, Natalie Jäger, Marc Remke, et al. 2012. “Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations” 148: 59–71.

Rezaee, Khosro, Gwanggil Jeon, Mohammad R. Khosravi, Hani H. Attar, and Alireza Sabzevari. 2022. “Deep Learning-Based Microarray Cancer Classification and Ensemble Gene Selection Approach.” *IET Systems Biology* 16 (3–4): 120–31.

Robert, Matius, and Karen Crasta. 2022. “Breaking the Vicious Circle: Extrachromosomal Circular DNA as an Emerging Player in Tumour Evolution.” *Seminars in Cell & Developmental Biology* 123 (March): 140–50.

Rosswog, Carolina, Christoph Bartenhagen, Anne Welte, Yvonne Kahlert, Nadine Hemstedt, Witali Lorenz, Maria Cartolano, et al. 2021. “Chromothripsis Followed by Circular Recombination Drives Oncogene Amplification in Human Cancer.” *Nature Genetics* 53 (12): 1673–85.

Sabath, D. E. 2013. “Philadelphia Chromosome.” In *Brenner’s Encyclopedia of Genetics (Second Edition)*, edited by Stanley Maloy and Kelly Hughes, 308. San Diego: Academic Press.

Saglam, Ozlen, Yin Xiong, Douglas C. Marchion, Carolina Strosberg, Robert M. Wenham, Joseph J. Johnson, Daryoush Saeed-Vafa, Christopher Cubitt, Ardeshir Hakam, and Anthony M. Magliocco. 2017. “ERBB4 Expression in Ovarian Serous Carcinoma Resistant to Platinum-Based Therapy.” *Cancer Control: Journal of the Moffitt Cancer Center* 24 (1): 89–95.

Sakamoto, Yoshitaka, Shuhei Miyake, Miho Oka, Akinori Kanai, Yosuke Kawai, Satoi Nagasawa, Yuichi Shiraishi, et al. 2022. “Phasing Analysis of Lung Cancer Genomes Using a Long Read Sequencer.” *Nature Communications* 13 (1): 1–17.

Sansregret, Laurent, and Charles Swanton. 2017. “The Role of Aneuploidy in Cancer Evolution.” *Cold Spring Harbor Perspectives in Medicine* 7 (1). https://doi.org/10.1101/cshperspect.a028373.

Sapoznik, Stav, Sarit Aviel-Ronen, Keren Bahar-Shany, Oranit Zadok, and Keren Levanon. 2017. “CCNE1 Expression in High Grade Serous Carcinoma Does Not Correlate with Chemoresistance.” *Oncotarget* 8 (37): 62240–47.

Schwab, M., K. Alitalo, K. H. Klempnauer, H. E. Varmus, J. M. Bishop, F. Gilbert, G. Brodeur, M. Goldstein, and J. Trent. 1983. “Amplified DNA with Limited Homology to Myc Cellular Oncogene Is Shared by Human Neuroblastoma Cell Lines and a Neuroblastoma Tumour.” *Nature* 305 (5931): 245–48.

Schwarze, Katharina, James Buchanan, Jenny C. Taylor, and Sarah Wordsworth. 2018. “Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (10): 1122–30.

Shah, Parth, Anil Aktas-Samur, Mariateresa Fulciniti, Raphael Szalat, Masood A. Shammas, Paul G. Richardson, Florence Magrangeas, et al. 2021. “Presence of Extrachromosomal DNA (EcDNA) Impacts Both Progression Free and Overall Survival and Is an Independent Poor Prognostic Marker in Multiple Myeloma.” *Blood* 138 (Supplement 1): 461–461.

Shao, Xin, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. 2019. “Copy Number Variation Is Highly Correlated with Differential Gene Expression: A Pan-Cancer Study.” *BMC Medical Genetics* 20 (1): 175.

Shen, Michael M. 2013. “Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome.” *Cancer Cell*.

Shen, Ronglai, and Venkatraman E. Seshan. 2016. “FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool for High-Throughput DNA Sequencing.” *Nucleic Acids Research* 44 (16): 1–9.

Shih, Andrew J., Andrew Menzin, Jill Whyte, John Lovecchio, Anthony Liew, Houman Khalili, Tawfiqul Bhuiya, Peter K. Gregersen, and Annette T. Lee. 2018. "Identification of Grade and Origin Specific Cell Populations in Serous Epithelial Ovarian Cancer by Single Cell RNA-Seq." *PloS One* 13 (11): e0206785.

Shih, Ie-Ming, and Robert J. Kurman. 2004. "Ovarian Tumorigenesis: A Proposed Model Based on Morphological and Molecular Genetic Analysis." *The American Journal of Pathology* 164 (5): 1511–18.

Shlien, Adam, and David Malkin. 2009. "Copy Number Variations and Cancer." *Genome Medicine* 1 (6): 62.

Shorokhova, Mariia, Nikolay Nikolsky, and Tatiana Grinchuk. 2021. "Chromothripsis-Explosion in Genetic Science." *Cells* 10 (5). https://doi.org/10.3390/cells10051102.

Shoshani, Ofer, Simon F. Brunner, Rona Yaeger, Peter Ly, Yael Nechemia-Arbely, Dong Hyun Kim, Rongxin Fang, et al. 2020. "Chromothripsis Drives the Evolution of Gene Amplification in Cancer." *Nature* 591 (December 2018). https://doi.org/10.1038/s41586-020-03064-z.

Singh, Shiva. 2018. "The Hundred-Dollar Genome: A Health Care Cart before the Genomic Horse." *CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*.

Skuja, Elina, Dagnija Kalniete, Miki Nakazawa-Miklasevica, Zanda Daneberga, Arnis Abolins, Gunta Purkalne, and Edvins Miklasevics. 2017. "Chromothripsis and Progression-Free Survival in Metastatic Colorectal Cancer." *Molecular and Clinical Oncology* 6 (2): 182–86.

Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers." *Nature Reviews. Cancer* 18 (11): 696–705.

Souche, Erika, Sergi Beltran, Erwin Brosens, John W. Belmont, Magdalena Fossum, Olaf Riess, Christian Gilissen, et al. 2022. "Recommendations for Whole Genome Sequencing in Diagnostics for Rare Diseases." *European Journal of Human Genetics: EJHG*, May, 1–5.

Steele, Christopher D., Ammal Abbasi, S. M. Ashiqul Islam, Amy L. Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, et al. 2022. “Signatures of Copy Number Alterations in Human Cancer.” *Nature* 606 (7916): 984–91.

Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, et al. 2011. “Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development.” *Cell* 144 (1): 27–40.

Sugita, Itsuki, Shohei Matsuyama, Hiroki Dobashi, Daisuke Komura, and Shumpei Ishikawa. 2022. “Viola: A Structural Variant Signature Extractor with User-Defined Classifications.” *Bioinformatics* 38 (2): 540–42.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. “Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering.” *Bioinformatics* 22 (12): 1540–42.

Tai, Chun-San, Yan-Ren Lin, Tsung-Han Teng, Ping-Yi Lin, Siang-Jyun Tu, Chih-Hung Chou, Ya-Rong Huang, et al. 2017. “Haptoglobin Expression Correlates with Tumor Differentiation and Five-Year Overall Survival Rate in Hepatocellular Carcinoma.” *PloS One* 12 (2): e0171269.

Takaya, Hisamitsu, Hidekatsu Nakai, Shiro Takamatsu, Masaki Mandai, and Noriomi Matsumura. 2020. “Homologous Recombination Deficiency Status-Based Classification of High-Grade Serous Ovarian Carcinoma.” *Scientific Reports* 10 (1): 1–8.

Talevich, Eric, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. 2016. “CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing.” *PLoS Computational Biology* 12 (4): 1–18.

Talseth-Palmer, Bente A., and Rodney J. Scott. 2011. “Genetic Variation and Its Role in Malignancy.” *International Journal of Biomedical Science: IJBS* 7 (3): 158–71.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (organisiation/Ins, Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, et al. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93.

Thouvenin, Laure, Mélinda Charrier, Sophie Clement, Yann Christinat, Jean-Christophe Tille, Mauro Frigeri, Krisztian Homicsko, et al. 2021. “Ovarian Cancer with High-

Level Focal ERBB2 Amplification Responds to Trastuzumab and Pertuzumab." *Gynecologic Oncology Reports* 37 (August): 100787.

Thul, Peter J., and Cecilia Lindskog. 2018. "The Human Protein Atlas: A Spatial Map of the Human Proteome." *Protein Science: A Publication of the Protein Society* 27 (1): 233–44.

Turner, Kristen M., Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, et al. 2017. "Extrachromosomal Oncogene Amplification Drives Tumour Evolution and Genetic Heterogeneity." *Nature* 543 (7643): 122–25.

Umbreit, Neil T., Cheng-Zhong Zhang, Luke D. Lynch, Logan J. Blaine, M. Anna, Richard Tourdot, Lili Sun, et al. 2019. "MECHANISMS GENERATING CANCER GENOME COMPLEXITY."

Verhaak, Roel G. W., Vineet Bafna, and Paul S. Mischel. 2019. "Extrachromosomal Oncogene Amplification in Tumour Pathogenesis and Evolution." *Nature Reviews. Cancer* 19 (5): 283–88.

Vinceti, Alessandro, Emre Karakoc, Clare Pacini, Umberto Perron, Riccardo Roberto De Lucia, Mathew J. Garnett, and Francesco Iorio. 2021. "CoRe: A Robustly Benchmarked R Package for Identifying Core-Fitness Genes in Genome-Wide Pooled CRISPR-Cas9 Screens." *BioRxiv*. https://doi.org/10.1101/2021.05.25.445610.

Voronina, Natalia, John K. L. Wong, Daniel Hübschmann, Mario Hlevnjak, Sebastian Uhrig, Christoph E. Heilig, Peter Horak, et al. 2020. "The Landscape of Chromothripsis across Adult Cancer Types." *Nature Communications* 11 (1): 1–13.

Wang, Yi Kan, Ali Bashashati, Michael S. Anglesio, Dawn R. Cochrane, Diljot S. Grewal, Gavin Ha, Andrew McPherson, et al. 2017. "Genomic Consequences of Aberrant DNA Repair Mechanisms Stratify Ovarian Cancer Histotypes." *Nature Genetics* 49 (6): 856–64.

Waszak, Sebastian M., Paul A. Northcott, Ivo Buchhalter, Giles W. Robinson, Christian Sutter, Susanne Groebner, Kerstin B. Grund, et al. 2018. "Spectrum and Prevalence of Genetic Predisposition in Medulloblastoma: A Retrospective

Genetic Study and Prospective Validation in a Clinical Trial Cohort." *The Lancet Oncology* 19 (6): 785–98.

Weber, Lea, Désirée Maßberg, Christian Becker, Janine Altmüller, Burkhard Ubrig, Gabriele Bonatz, Gerhard Wölk, et al. 2018. "Olfactory Receptors as Biomarkers in Human Breast Carcinoma Tissues." *Frontiers in Oncology* 8 (February): 33.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." *Nature Genetics* 45 (10): 1113–20.

Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O. Korbel. 2013. "Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease." *Nature Reviews. Genetics* 14 (2): 125–38.

Yamulla, Robert Joseph, Shreya Nalubola, Andrea Flesken-Nikitin, Alexander Yu Nikitin, and John C. Schimenti. 2020. "Most Commonly Mutated Genes in High-Grade Serous Ovarian Carcinoma Are Nonessential for Ovarian Surface Epithelial Stem Cell Transformation." *Cell Reports* 32 (9): 108086.

Yi, Eunhee, Amit D. Gujar, Molly Guthrie, Hoon Kim, Dacheng Zhao, Kevin C. Johnson, Samirkumar B. Amin, et al. 2022. "Live-Cell Imaging Shows Uneven Segregation of Extrachromosomal DNA Elements and Transcriptionally Active Extrachromosomal DNA Hubs in Cancer." *Cancer Discovery* 12 (2): 468–83.

Yu, Zhenhua, Yuanning Liu, Yi Shen, Minghui Wang, and Ao Li. 2014. "CLImAT: Accurate Detection of Copy Number Alteration and Loss of Heterozygosity in Impure and Aneuploid Tumor Samples Using Whole-Genome Sequencing Data." *Bioinformatics* 30 (18): 2576–83.

Zakov, Shay, Marcus Kinsella, and Vineet Bafna. 2013. "An Algorithmic Approach for Breakage-Fusion-Bridge Detection in Tumor Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): 5546–51.

Zeng, Xixi, Maoping Wan, and Jianmin Wu. 2020. "EcDNA within Tumors: A New Mechanism That Drives Tumor Heterogeneity and Drug Resistance." *Signal Transduction and Targeted Therapy* 5 (1): 3–4.

Zhang, Cheng Zhong, Alexander Spektor, Hauke Cornils, Joshua M. Francis, Emily K.

Jackson, Shiwei Liu, Matthew Meyerson, and David Pellman. 2015. "Chromothripsis from DNA Damage in Micronuclei." *Nature* 522 (7555): 179–84.

Zhang, Xiaohong, Omar De la Cruz, Jayant M. Pinto, Dan Nicolae, Stuart Firestein, and Yoav Gilad. 2007. "Characterizing the Expression of the Human Olfactory Receptor Gene Family Using a Novel DNA Microarray." *Genome Biology* 8 (5): R86.

Zhao, Min, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. 2013. "Computational Tools for Copy Number Variation (CNV) Detection Using next-Generation Sequencing Data: Features and Perspectives." *BMC Bioinformatics* 14 Suppl 11 (September): S1.

Zhao, Xue-Ke, Pengwei Xing, Xin Song, Miao Zhao, Linxuan Zhao, Yonglong Dang, Ling-Ling Lei, et al. 2021a. "Extrachromosomal DNA Is Associated with Chromothripsis Events and Diverse Prognoses in Gastric Cardia Adenocarcinoma." *BioRxiv*. https://doi.org/10.1101/2021.07.02.450861.

———. 2021b. "Focal Amplifications Are Associated with Chromothripsis Events and Diverse Prognoses in Gastric Cardia Adenocarcinoma." *Nature Communications* 12 (1): 1–14.

Zhu, Yanfen, Amit D. Gujar, Chee-Hong Wong, Harianto Tjong, Chew Yee Ngan, Liang Gong, Yi-An Chen, et al. 2021. "Oncogenic Extrachromosomal DNA Functions as Mobile Enhancers to Globally Amplify Chromosomal Transcription." *Cancer Cell* 39 (5): 694-707.e7.

Appendix

Overlap of CCNE1 and BFB

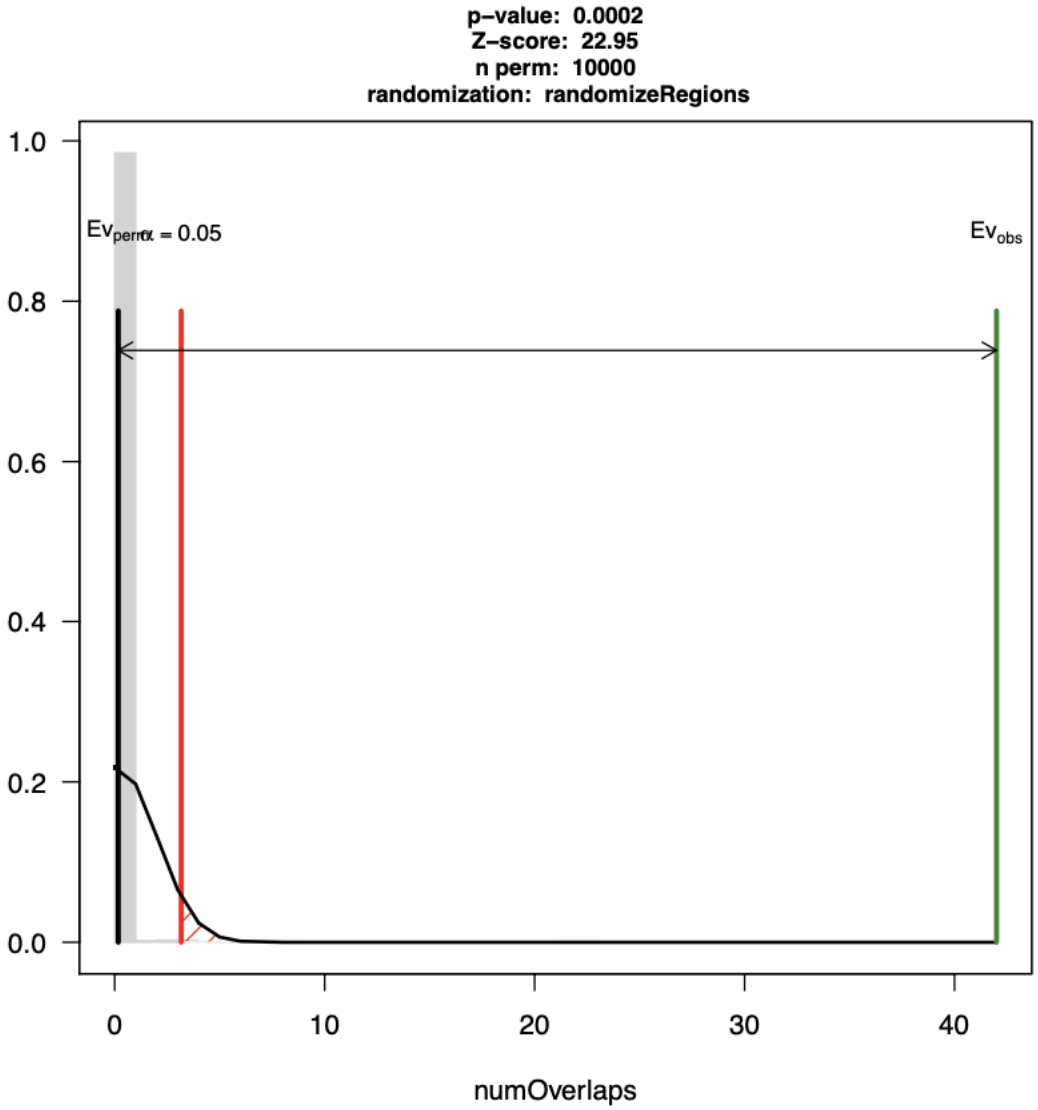
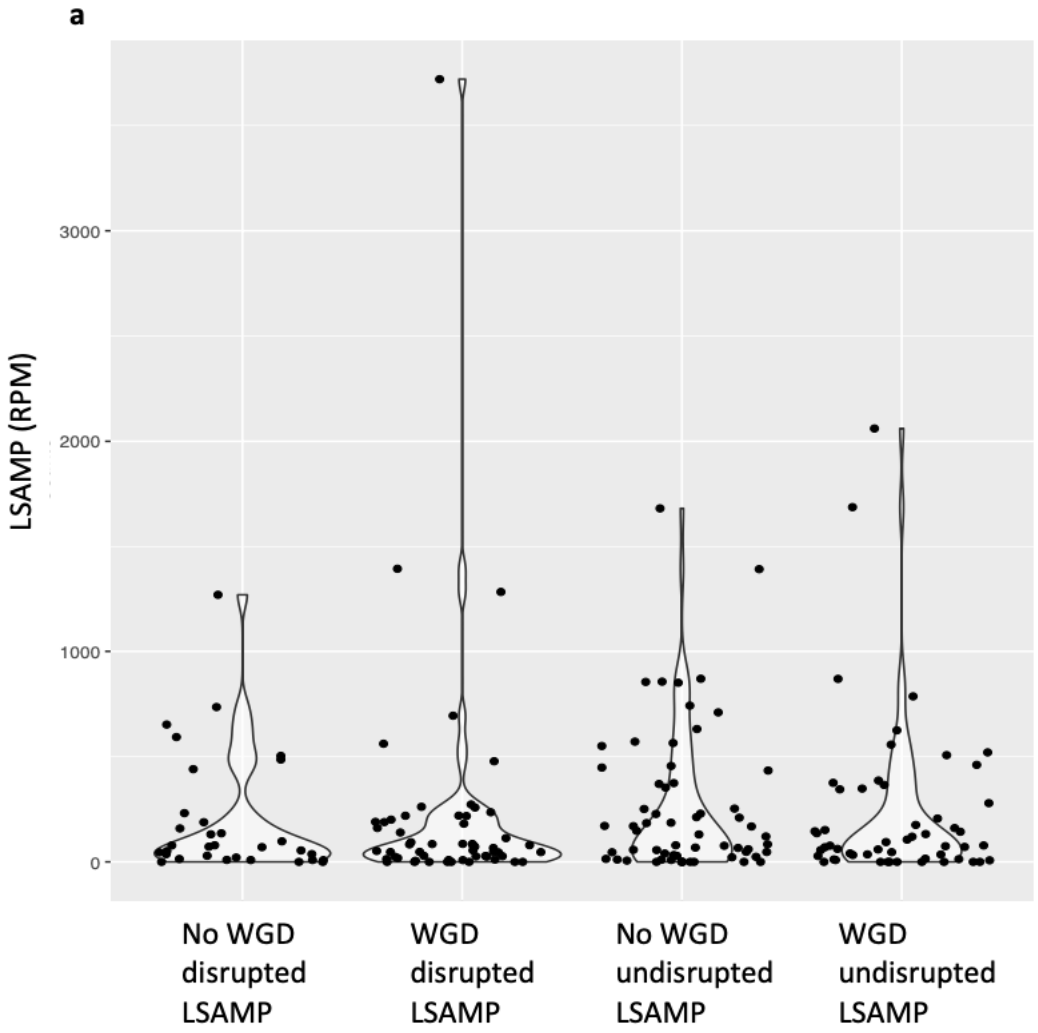


Figure 69 Overlap of CCNE1 and BFB

The result of circular permutation between BFB regions and the CCNE1 gene.



b

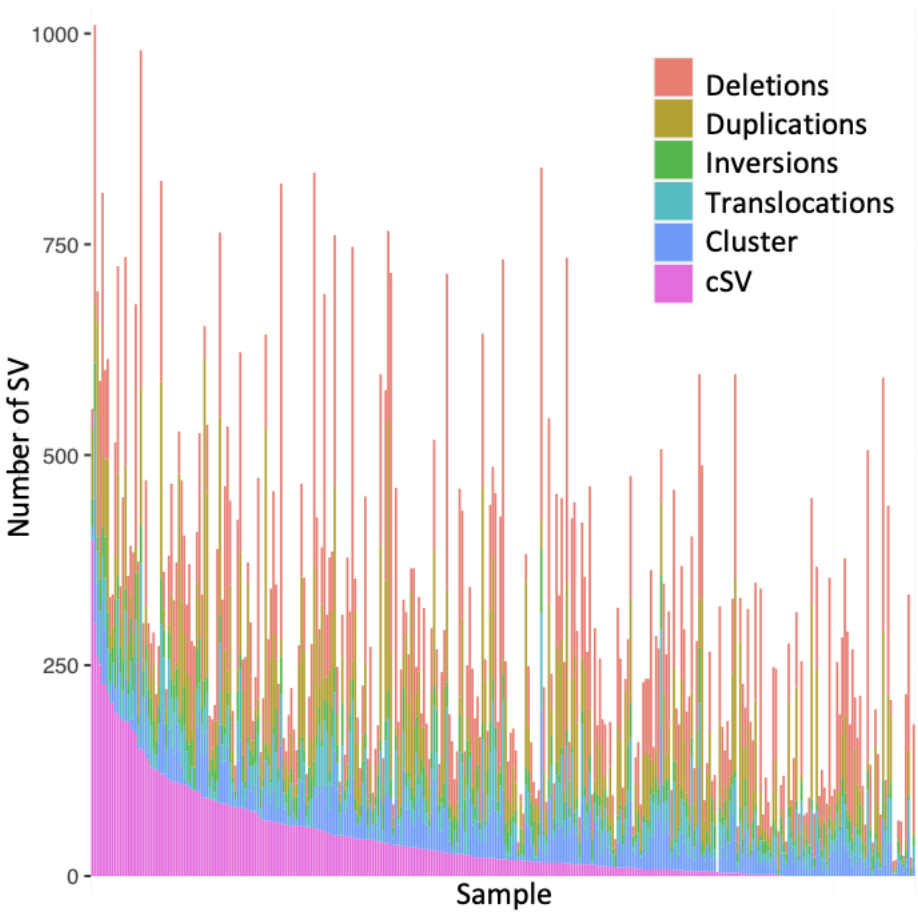
	Median	IQR
Samples without WGD and disrupted LSAMP	71.09	180.29
Samples with WGD and disrupted LSAMP	69.01	170.09
Samples without WGD and without disrupted LSAMP	125.43	344.27
Samples with WGD and without disrupted LSAMP	77.51	309.9

Figure 70 Impact on LSAMP Gene Expression of disruption and whole genome doubling

The normalised reads per million for LSAMP split in to four groups sample with disrupted LSAMP but without whole genome doubling (WGD), samples with WGD and disrupted LSAMP, samples without disrupted LSAMP or WGD, and samples with WGD and undisrupted (**a**). summary statistics for median and interquartile range (**b**). A Wilcoxon test was used to test if there were any significant difference in expression between group and none were significant.

Structural Variants Simple and clustered

a



b

SV Classification	Number of SVs
cSV	13877
Cluster	10294
Dup	20716
Del	33296
Tra	6600
Inv	10517

Figure 71 Structural variant classification

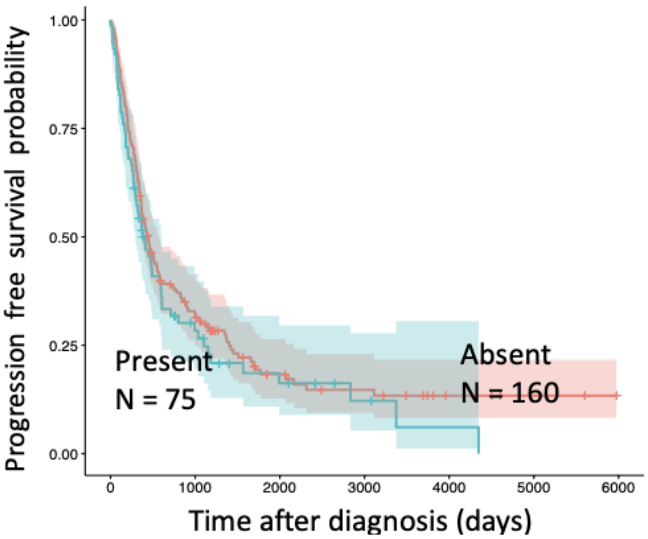
across the combined cohort

The number of SV in each sample of the combined cohort classified into seven groups; clustered SV where at least 50% of breakpoints are within a cSV region (pink), clustered SV where at least 50% of breakpoints not are within a cSV region (purple), un-clustered insertions (blue), un-clustered Inversions (teal), un-clustered translocations (green), un-clustered duplications (gold), and un-clustered deletions (red). The samples were

ordered from samples with the greatest umber SV explained by cSV to fewest. The number of SV for each classification is shown in a table (**b**)

Progression free survival of cSVs

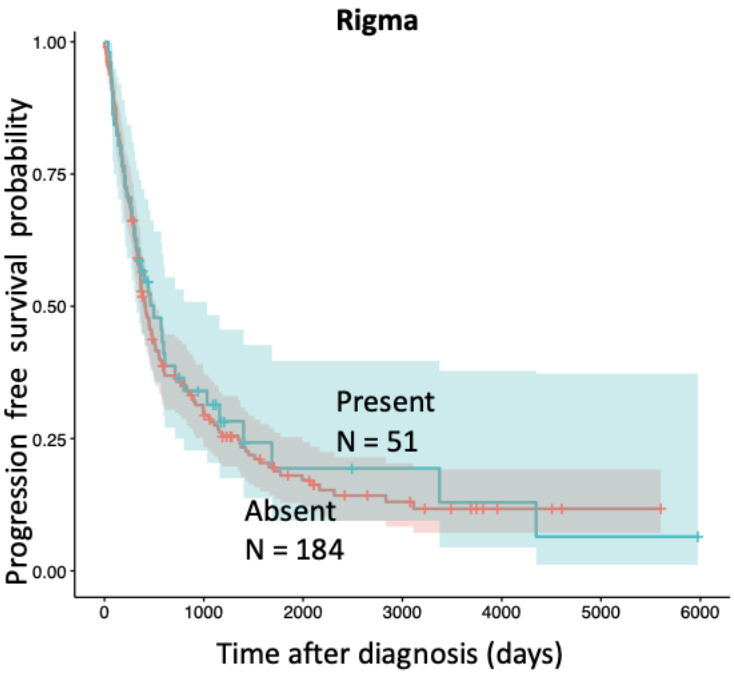
a i



a ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99-1.02)	0.421	
Stage	N=235	1.34 (1.0.6-1.71)	0.015	
HRD	Absent N=90	Reference		
	Present N=145	1.24 (0.42-0.81)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.24 (0.93-1.67)	0.145	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.20 (2.19-4.68)	0.001	
	BCCA N=58	1.40 (0.94-2.09)	0.097	
	TCGA N=0			
Chromothripsis	Absent N=160	Reference		
	Present N=75	0.96 (0.69-1.34)	0.816	

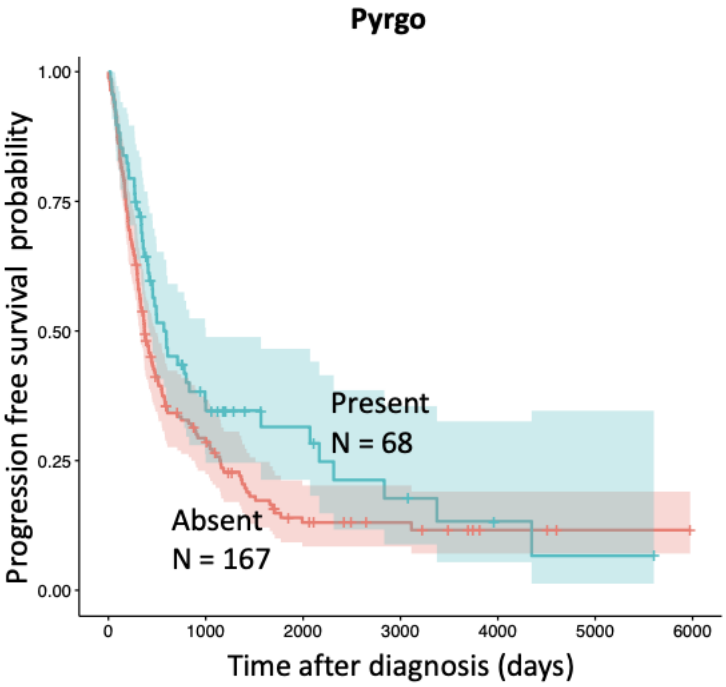
b i



b ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99-1.02)	0.464	
Stage	N=235	1.34 (1.07-1.74)	0.011	
HRD	Absent N=90	Reference		
	Present N=145	0.58 (0.42-0.81)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.24 (0.93-1.66)	0.145	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.24 (2.22-4.72)	0.001	
	BCCA N=58	1.44 (0.97-2.15)	0.07	
	TCGA N=0			
Rigma	Absent N=181	Reference		
	Present N=51	1.17 (0.81-1.69)	0.408	

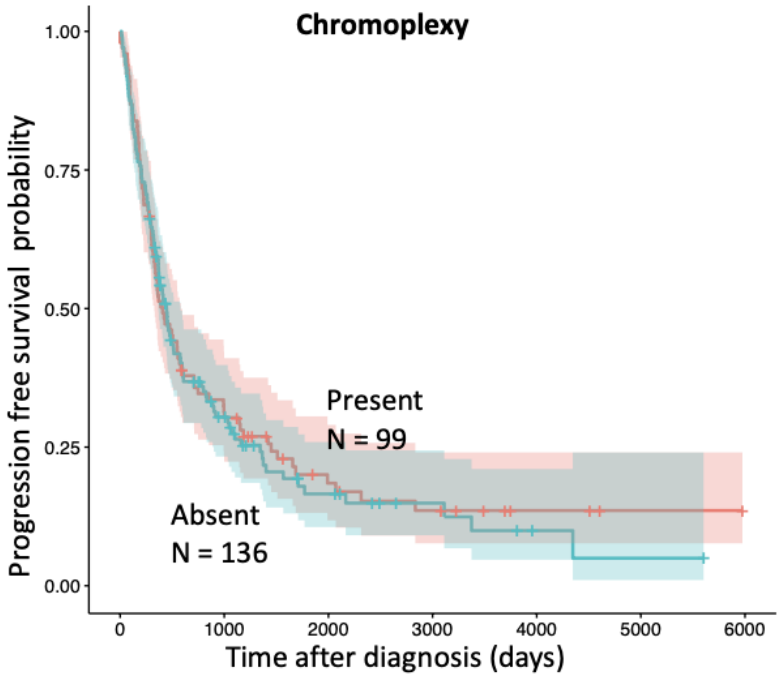
c i



c ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99-1.02)	0.333	
Stage	N=235	1.33 (1.05-1.69)	0.02	
HRD	Absent N=90	Reference		
	Present N=145	0.56 (0.40-0.78)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.18 (0.88-1.59)	0.275	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.21 (2.20-4.67)	0.001	
	BCCA N=58	1.42 (0.96-2.11)	0.078	
	TCGA N=0			
Pyrgo	Absent N=167	Reference		
	Present N=68	0.79 (0.56-1.13)	0.196	

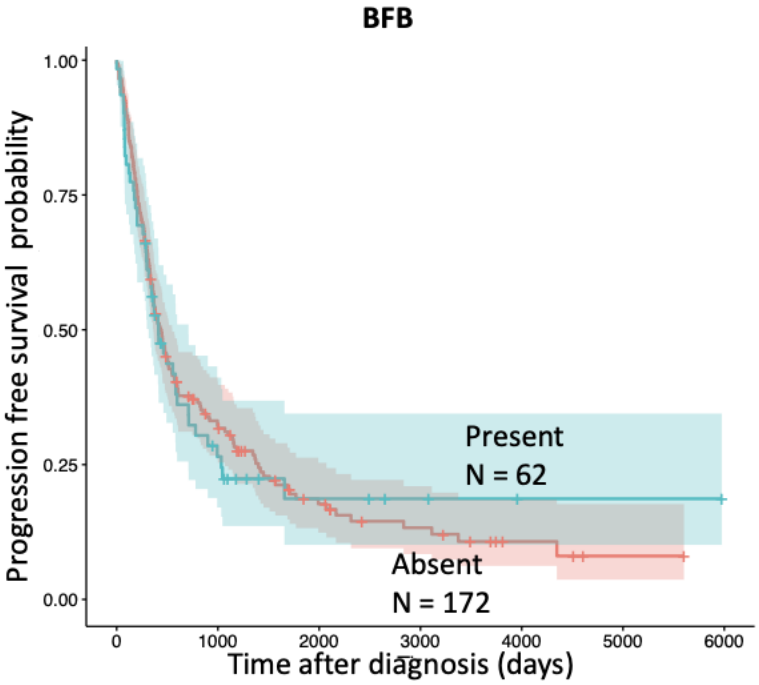
d i



d ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99-1.02)	0.460	
Stage	N=235	1.35 (1.06-1.71)	0.016	
HRD	Absent N=90	Reference		
	Present N=145	0.58 (0.42-0.81)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.23 (0.91-1.66)	0.185	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.19 (2.19-4.64)	0.001	
	BCCA N=58	1.42 (0.96-2.11)	0.082	
	TCGA N=0			
Chromoplexy	Absent N=99	Reference		
	Present N=136	1.04 (0.76-1.43)	0.792	

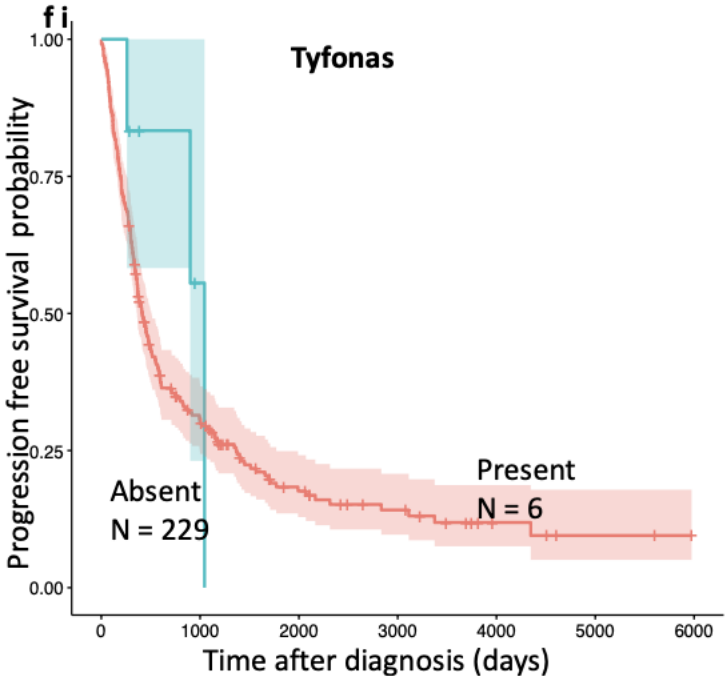
e i



e ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99 -1.02)	0.365	
Stage	N=235	1.35 (1.07-1.72)	0.013	
HRD	Absent N=90	Reference		
	Present N=145	0.56 (0.40-0.79)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.24 (0.93-1.66)	0.146	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.19 (2.19-4.63)	0.001	
	BCCA N=58	1.41 (0.65-2.08)	0.089	
	TCGA N=0			
BFB	Absent N=173	Reference		
	Present N=62	0.86 (0.61-1.28)	0.523	

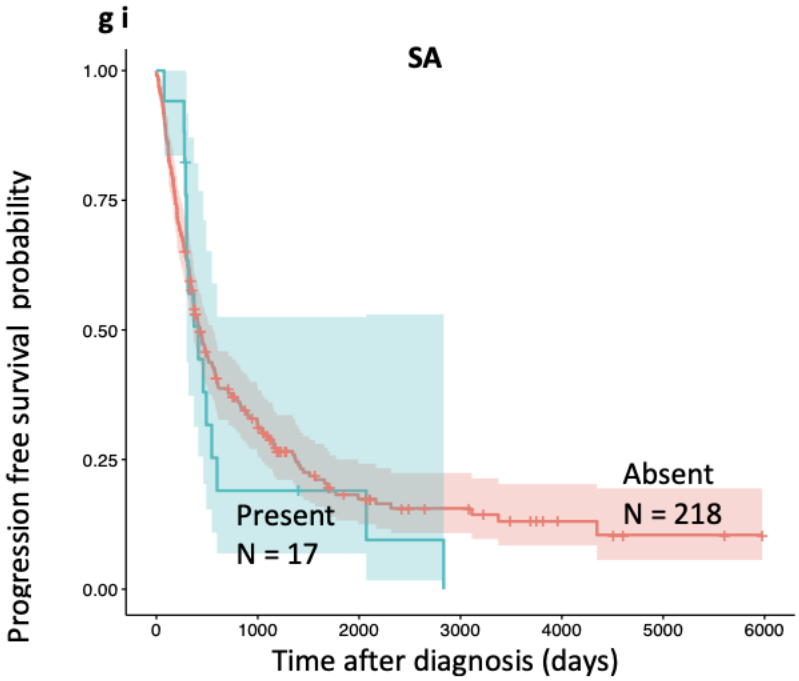
f i



f ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99 -1.02)	0.365	
Stage	N=235	1.37 (1.07-1.75)	0.011	
HRD	Absent N=90	Reference		
	Present N=145	0.57 (0.41-0.79)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.26 (0.94-1.69)	0.117	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.09 (2.13-4.49)	0.001	
	BCCA N=58	1.39 (0.94-2.05)	0.103	
	TCGA N=0			
Tyfonas	Absent N=229	Reference		
	Present N=6	0.53 (0.16-1.72)	0.289	

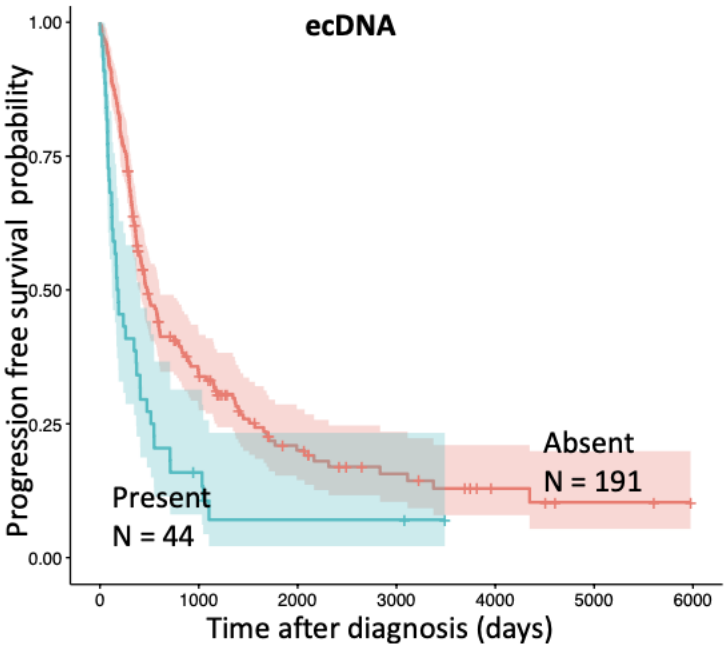
g i



g ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.01 (0.99 -1.02)	0.416	
Stage	N=235	1.35 (1.07-1.75)	0.013	
HRD	Absent N=90	Reference		
	Present N=145	0.59 (0.43-0.81)	0.001	
WGD	Absent N=117	Reference		
	Present N=118	1.28 (0.95-1.72)	0.098	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.37 (2.30-4.93)	0.001	
	BCCA N=58	1.45 (0.98-2.15)	0.63	
	TCGA N=0			
SA	Absent N=229	Reference		
	Present N=6	1.73 (1.00-2.99)	0.052	

h i



h ii

Condition		Hazard ratio	P value	Forest plot
Age	N=235	1.00 (0.99 -1.02)	0.439	
Stage	N=235	1.40 (1.06-1.72)	0.014	
HRD	Absent N=90	Reference		
	Present N=145	0.60 (0.43-0.83)	0.002	
WGD	Absent N=117	Reference		
	Present N=118	1.28 (0.95-1.72)	0.16	
Cohort	SHGSOC N=97	Reference		
	AOCS N=80	3.00 (2.04-4.54)	0.001	
	BCCA N=58	1.40 (0.95-2.09)	0.086	
	TCGA N=0			
ecDNA	Absent N=191	Reference		
	Present N=44	1.1 (0.76-1.66)	0.557	

Figure 72 Progression free survival time

For each of the samples with and without cSV of eight types a Kaplan-Meier survival curve for progression free survival time after diagnosis is part **i** for panels **a** to **h**. The eight cSV investigated are; chromothripsis (**a**), Rigma (**b**), Pyrgo(**c**), chromoplexy (**d**), BFB (**e**), tyfonas (**f**), seismic amplification (**g**), and ecDNA (**h**). A Cox proportion hazard model comparing impact on progression free survival time after diagnosis adjusting for age, stage at diagnosis, HRD status, and WGD a hazard ratio of 1 (no effect) is shown by a dashed line in part **ii** for panels **a** to **h**. Multiple testing correction by Benjamini and Hochberg was performed on all p values.

SNVs, Indels, Survival and cSV

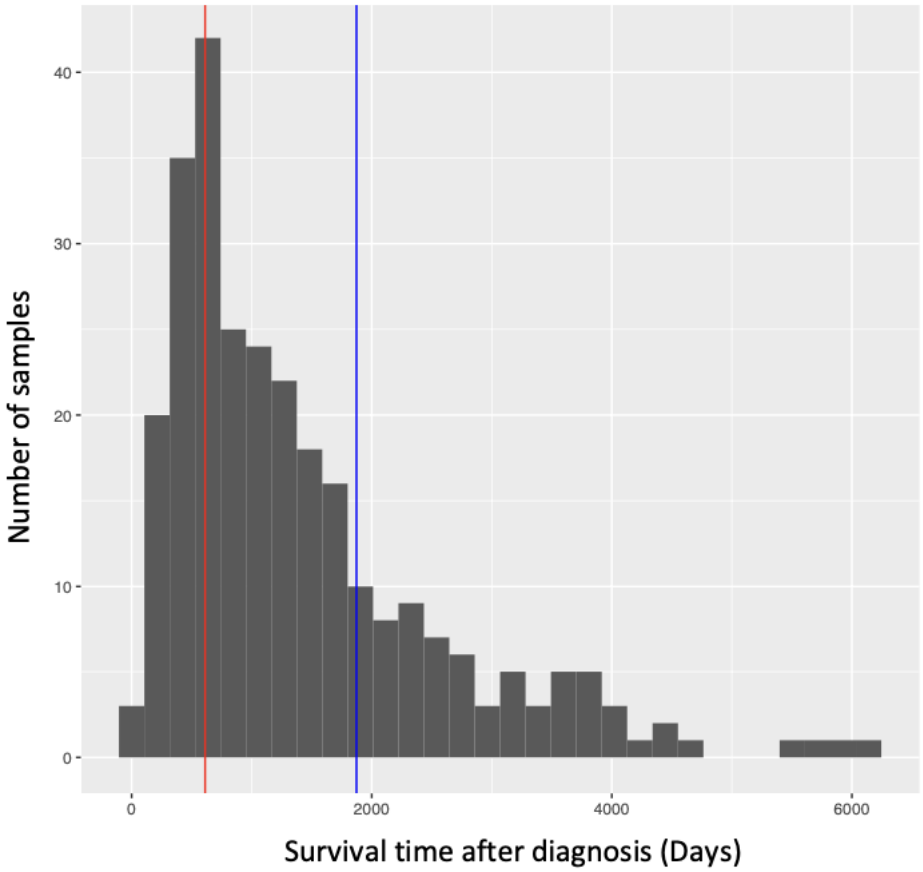


Figure 73 Survival after diagnosis

Histogram displaying the survival time after diagnosis. The red line show the cut off to be the 25% of samples with the shorted survival time. The blueline show the cut off to be the 25% of samples with the longest survival time.

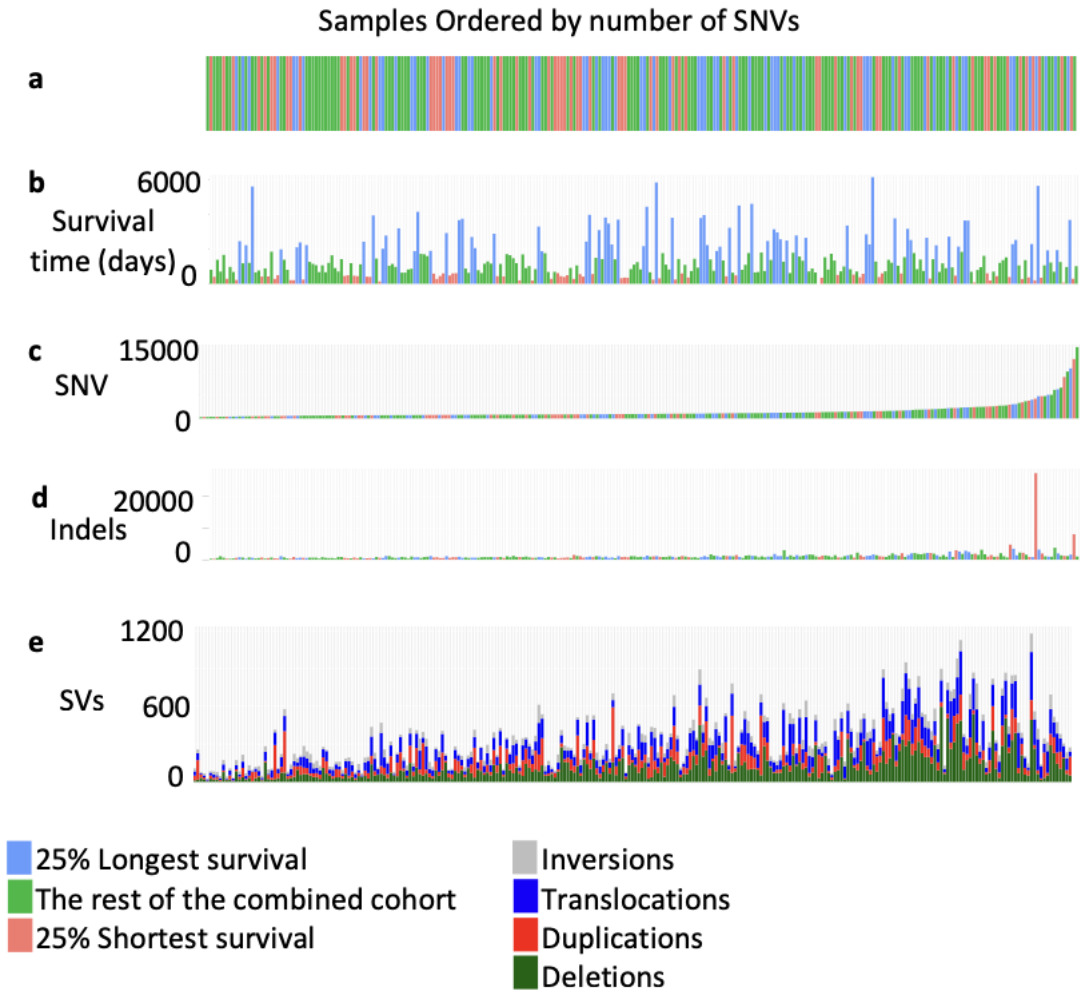


Figure 74 Survival, Indels, SVs and, SNVs ordered by SNV number

Samples ordered by number of single nucleotide variants (SNVs) with the greatest on the right coloured by survival group the top 25% longest surviving (Blue) the bottom 25% shortest surviving (red) and the rest of the combined cohort (Green) (**a**). The survival time after diagnosis in days of each sample (**b**). The number of SNV in each sample (**c**). The number of indels in each sample (**e**). The number of SNVs in a samples coloured by SNV type inversion (Gray), translocations (Blue), duplications (Red), Deletions (dark green).

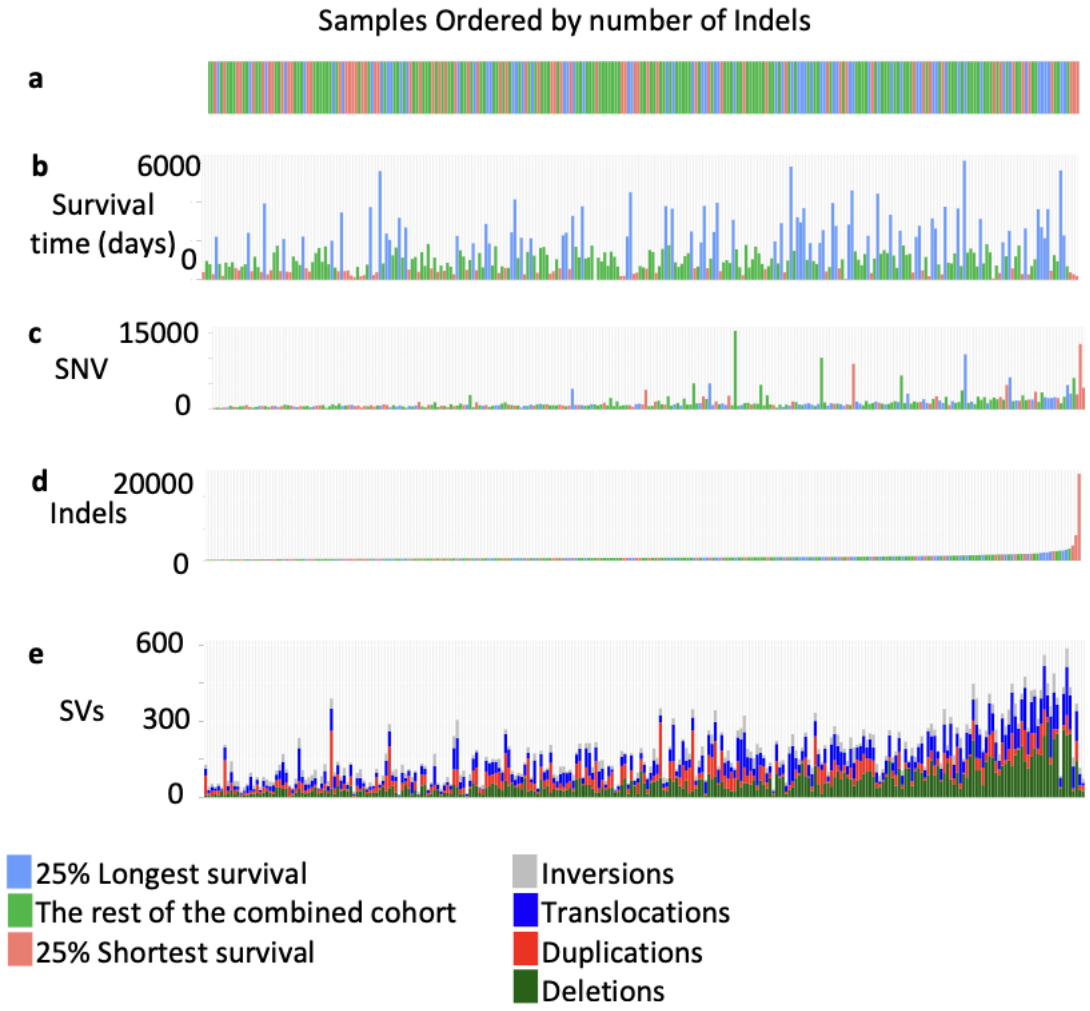


Figure 75 Survival, Indels, SVs and, SNVs ordered by Indels number

Samples ordered by number of indels in a sample with the greatest on the right coloured by survival group the top 25% longest surviving (Blue) the bottom 25% shortest surviving (red) and the rest of the combined cohort (Green) (**a**). The survival time after diagnosis in days of each sample (**b**). The number of SNV in each sample (**c**). The number of indels in each sample (**e**). The number of SNVs in a samples coloured by SNV type inversion (Gray), translocations (Blue), duplications (Red), Deletions (dark green).

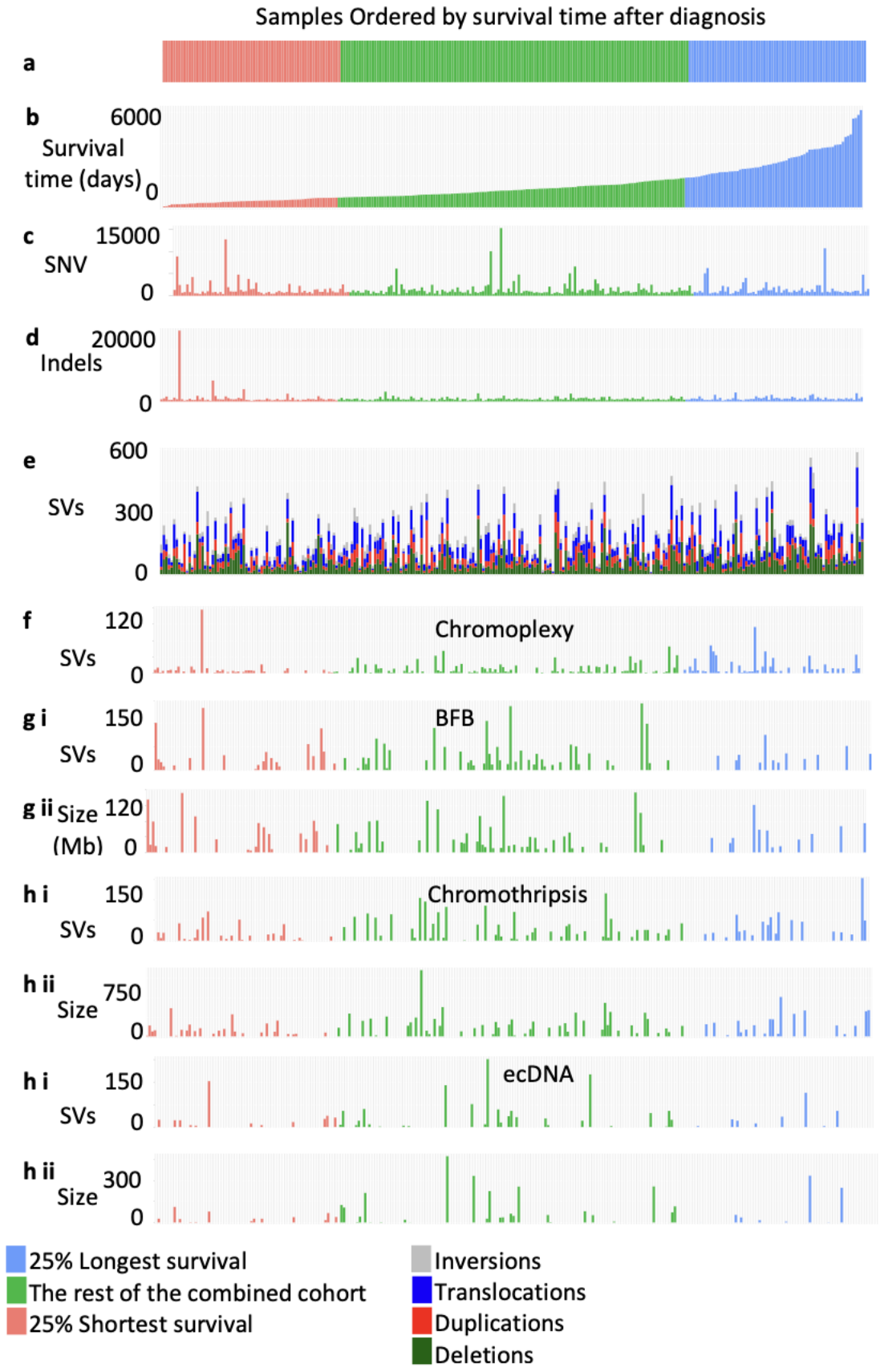
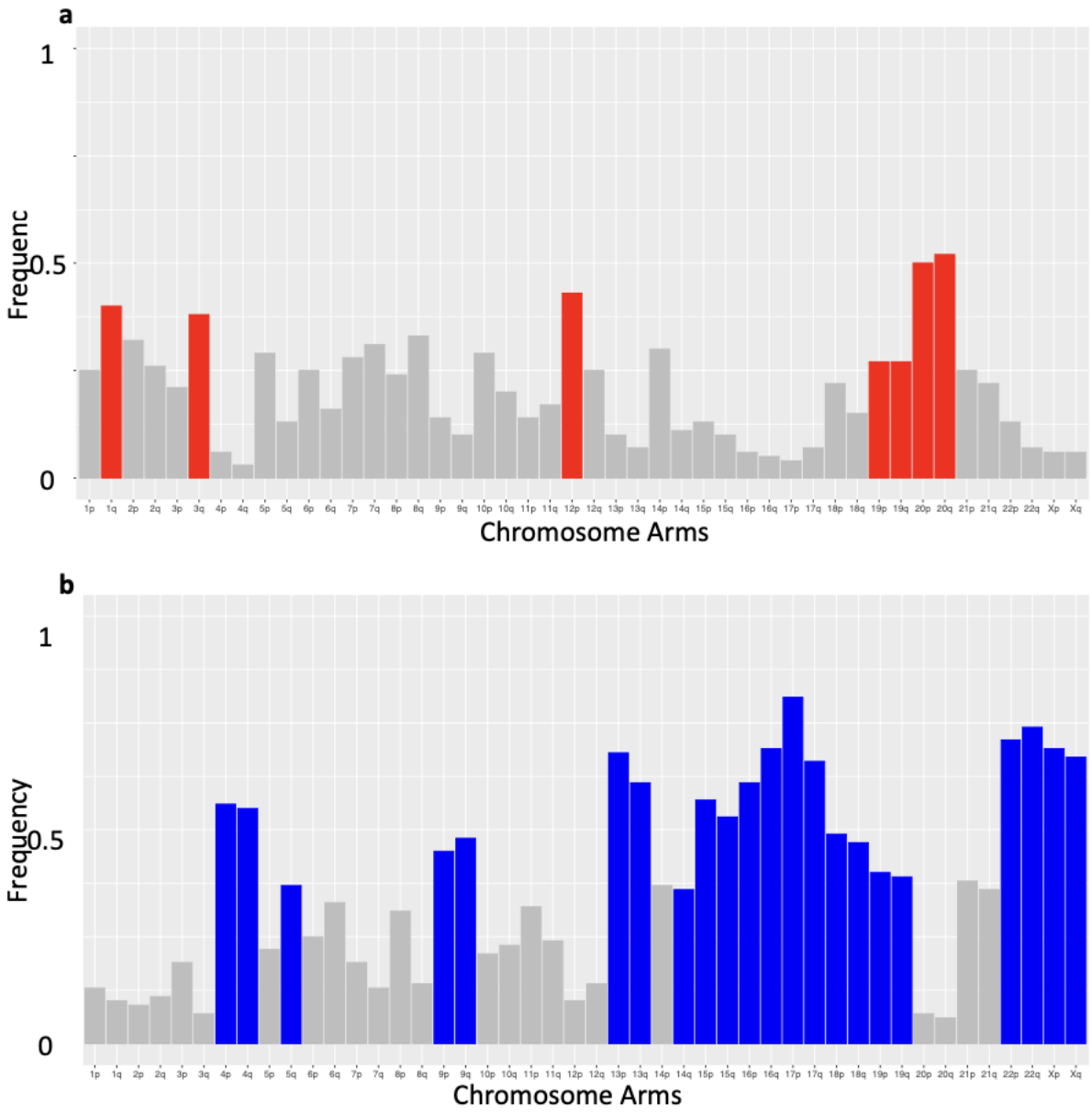


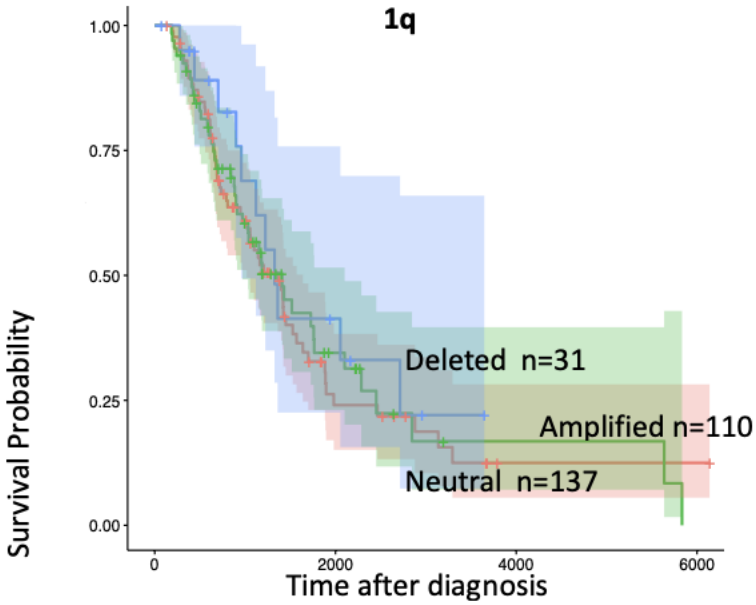
Figure 76 Survival, Indels, SVs and, SNVs ordered by Survival

Samples ordered by the survival time after diagnosis in a sample with the greatest on the right coloured by survival group the top 25% longest surviving (Blue) the bottom 25% shortest surviving (red) and the rest of the combined cohort (Green) (**a**). The survival time after diagnosis in days of each sample (**b**). The number of SNV in each sample (**c**). The number of indels in each sample (**e**). The number of SNVs in a samples coloured by SNV type inversion (Gray), translocations (Blue), duplications (Red), Deletions (dark green).

Chromosomal arm loss



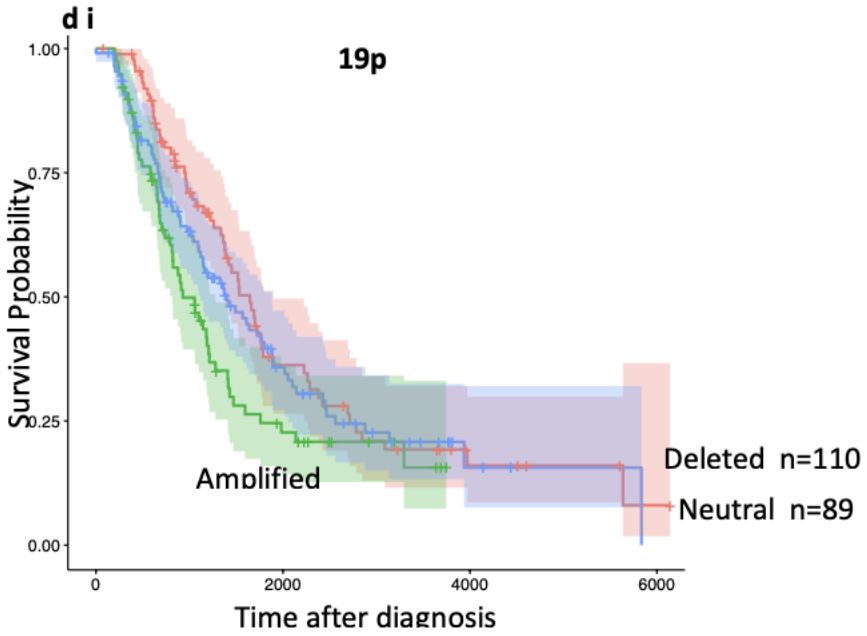
c i



c ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.01)	0.066	
Stage	N=278	1.48 (1.13-1.90)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.43 (0.31-0.60)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.86 (0.56-1.30)	0.94	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.74 (1.18-2.60)	0.005	
	BCCA N=59	0.86 (0.56-1.30)	0.49	
	TCGA N=31	1.68 (1.06-2.70)	0.027	
Arm 1q	Neutral N=137	Reference		
	Amplified N=110	0.79 (0.57-1.10)	0.141	
	Deletion N=31	0.71 (0.42-1.20)	0.197	

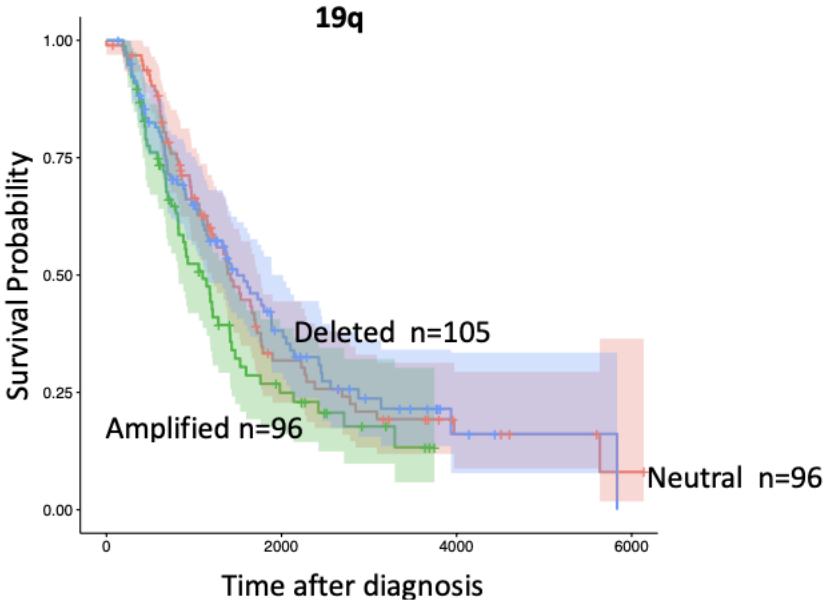
d i



d ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.098	
Stage	N=278	1.51 (1.14-1.99)	0.004	
HRD	Absent N=116	Reference		
	Present N=162	0.48 (0.35-0.65)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.70-1.27)	0.705	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.72 (1.26-2.56)	0.008	
	BCCA N=59	0.89 (0.58-1.37)	0.599	
	TCGA N=31	1.69 (1.07-2.68)	0.026	
Arm 19p	Neutral N=89	Reference		
	Amplified N=79	1.31 (0.89-1.95)	0.171	
	Deletion N=110	1.08 (0.75-1.56)	0.685	

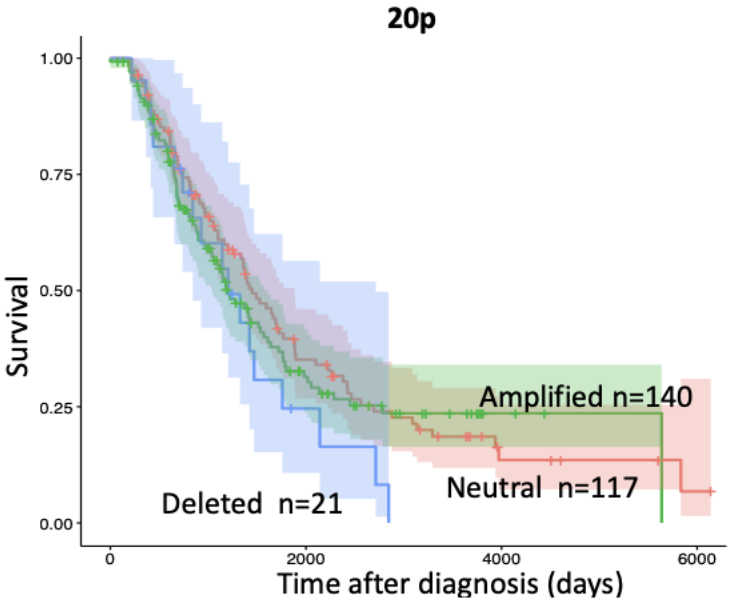
e i



e ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.093	
Stage	N=278	1.47 (1.11-1.94)	0.007	
HRD	Absent N=116	Reference		
	Present N=162	0.47 (0.34-0.64)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.71-1.29)	0.79	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.21-2.66)	0.004	
	BCCA N=59	0.868 (0.58-1.35)	0.569	
	TCGA N=31	1.70 (1.08-2.70)	0.023	
Arm 19q	Neutral N=96	Reference		
	Amplified N=77	1.16 (0.79-1.69)	0.452	
	Deletion N=105	0.94 (0.65-1.34)	0.718	

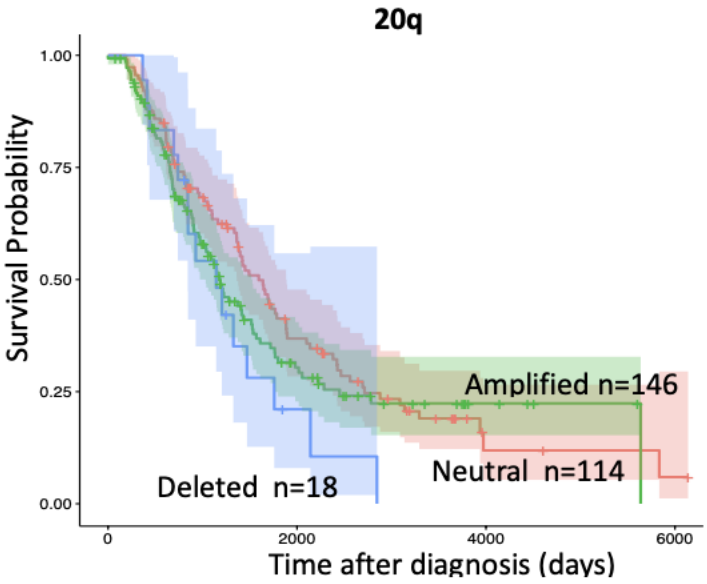
f i



f ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.055	
Stage	N=278	1.47 (1.12-1.94)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.97 (0.71-1.33)	0.858	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.21-2.63)	0.003	
	BCCA N=59	0.89 (0.58-1.39)	0.618	
	TCGA N=31	1.67 (1.04-2.68)	0.034	
Arm 20p	Neutral N=117	Reference		
	Amplified N=140	0.90 (0.64-1.28)	0.567	
	Deletion N=21	1.09 (0.62-1.93)	0.765	

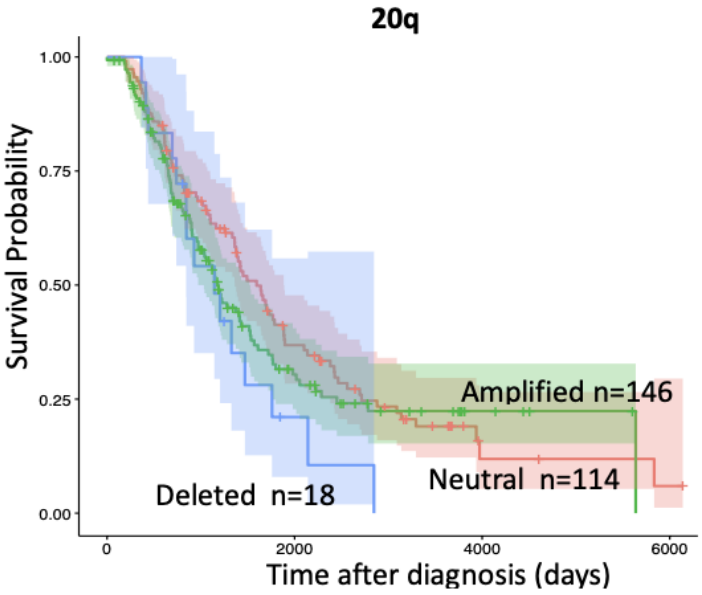
g i



g ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.076	
Stage	N=278	1.47 (1.12-1.94)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.47 (0.34-0.64)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.69-1.28)	0.678	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.80 (1.22-2.66)	0.003	
	BCCA N=59	0.91 (0.59-1.40)	0.656	
	TCGA N=31	1.70 (1.07-2.72)	0.026	
Arm 20q	Neutral N=114	Reference		
	Amplified N=146	1.04 (0.74-1.45)	0.829	
	Deletion N=18	1.18 (0.66-2.13)	0.578	

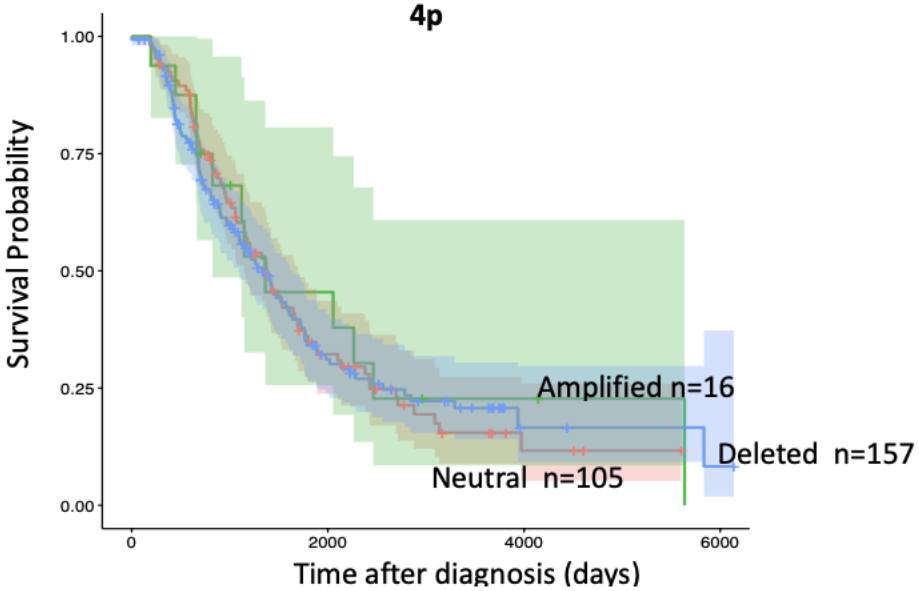
h i



h ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.076	
Stage	N=278	1.47 (1.12-1.94)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.47 (0.34-0.64)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.69-1.28)	0.678	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.80 (1.22-2.66)	0.003	
	BCCA N=59	0.91 (0.59-1.40)	0.656	
	TCGA N=31	1.70 (1.07-2.72)	0.026	
Arm 20q	Neutral N=114	Reference		
	Amplified N=146	1.04 (0.74-1.45)	0.829	
	Deletion N=18	1.18 (0.66-2.13)	0.578	

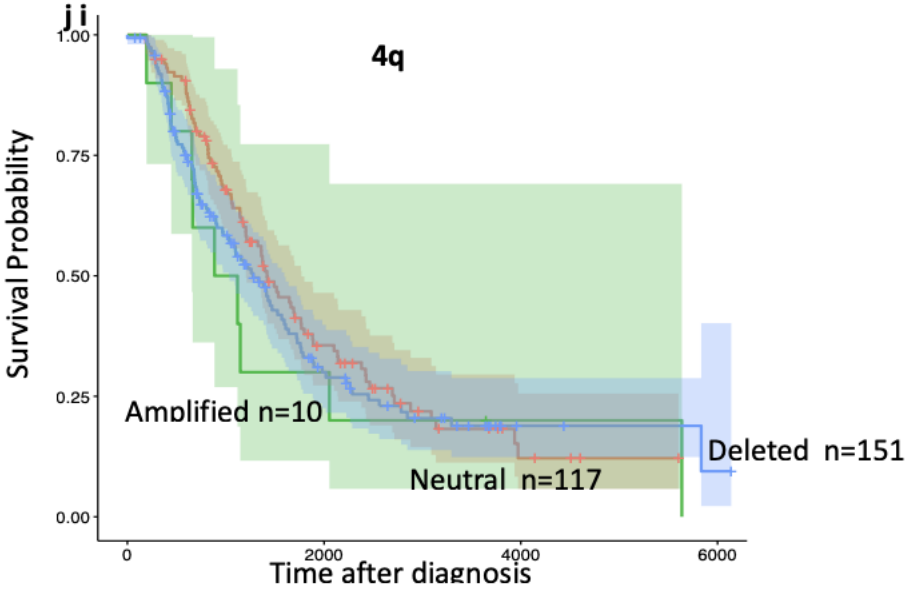
i i



i ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.051	
Stage	N=278	1.48 (1.13-1.95)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.33-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.72-1.30)	0.806	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.82 (1.24-2.69)	0.002	
	BCCA N=59	0.88 (0.57-1.35)	0.558	
	TCGA N=31	1.64 (1.04-2.60)	0.035	
Arm 4p	Neutral N=105	Reference		
	Amplified N=16	0.80 (0.43-1.50)	0.829	
	Deletion N=157	0.84 (0.61-1.14)	0.578	

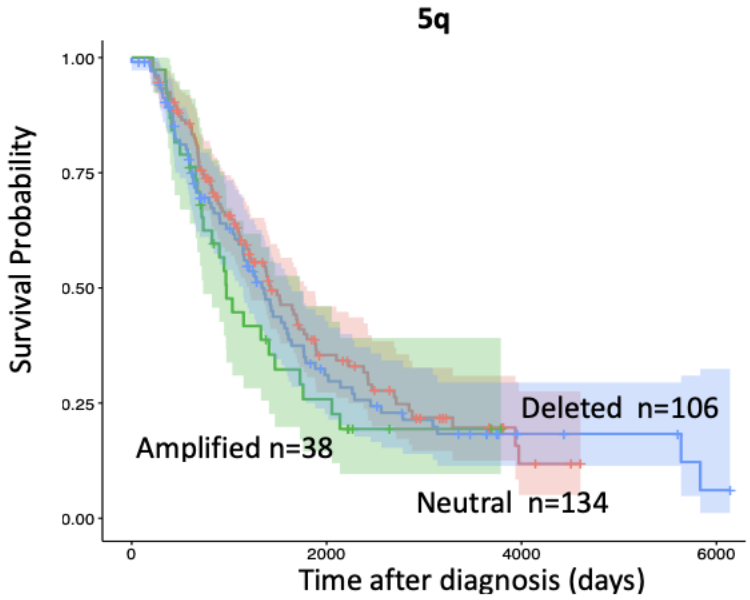
j i



j ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.065	
Stage	N=278	1.49 (1.13-1.95)	0.004	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.33-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.72-1.30)	0.725	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.82 (1.23-2.68)	0.003	
	BCCA N=59	0.90 (0.59-1.38)	0.633	
	TCGA N=31	1.67 (1.06-2.65)	0.028	
Arm 4q	Neutral N=117	Reference		
	Amplified N=10	1.12 (0.55-2.29)	0.762	
	Deletion N=151	0.84 (0.66-1.22)	0.485	

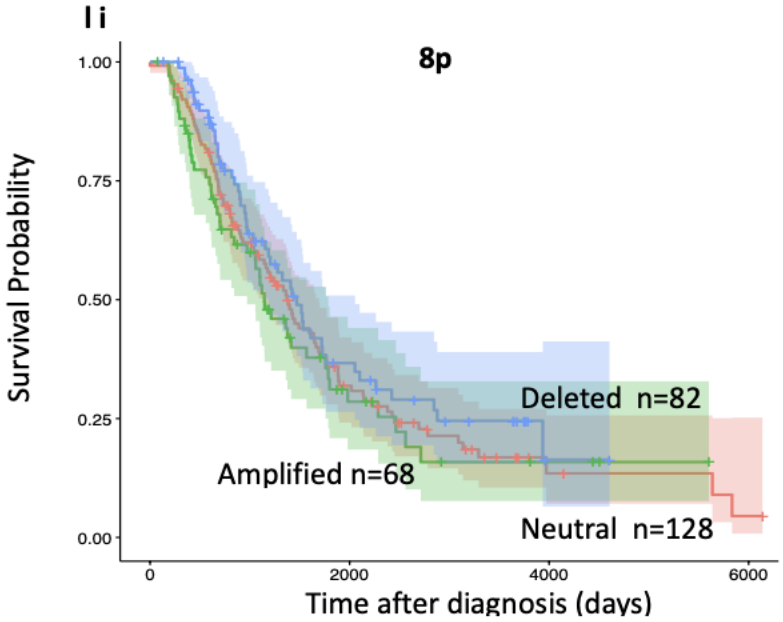
k i



k ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.073	
Stage	N=278	1.50 (1.13-1.99)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.69-1.27)	0.685	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.80 (1.21-2.67)	0.004	
	BCCA N=59	0.90 (0.58-1.38)	0.618	
	TCGA N=31	1.67 (1.06-2.65)	0.029	
Arm 5q	Neutral N=134	Reference		
	Amplified N=38	1.05 (0.66-1.67)	0.838	
	Deletion N=106	1.12 (0.81-1.55)	0.500	

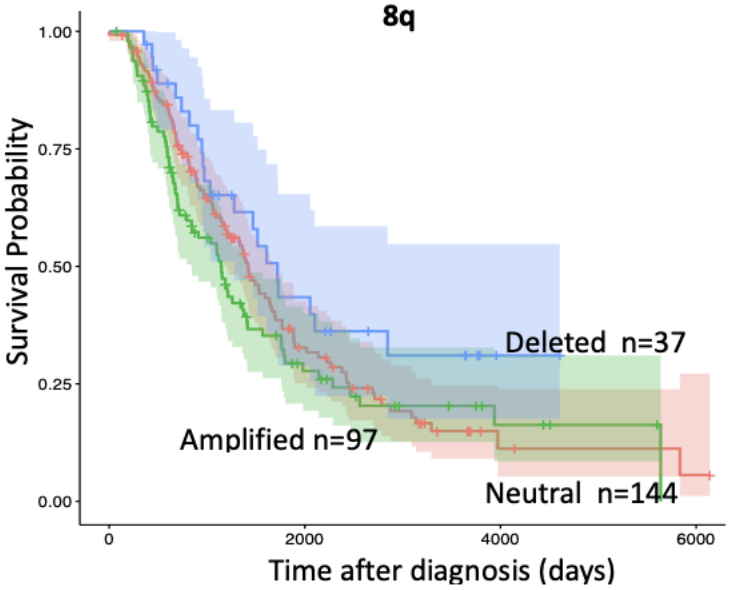
I i



I ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.078	
Stage	N=278	1.47 (1.12-1.93)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.34-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.70-1.30)	0.770	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.78 (1.21-2.64)	0.004	
	BCCA N=59	0.88 (0.57-1.36)	0.567	
	TCGA N=31	1.63 (1.02-2.60)	0.041	
Arm 8p	Neutral N=128	Reference		
	Amplified N=68	1.09 (0.74-1.59)	0.656	
	Deletion N=82	1.12 (0.58-1.19)	0.315	

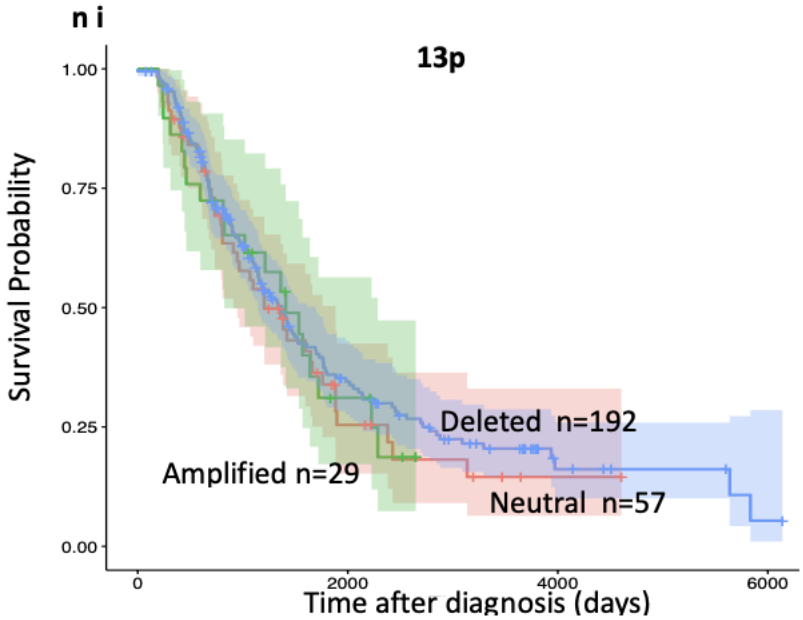
m i



m ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.080	
Stage	N=278	1.47 (1.12-1.93)	0.003	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.34-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.70-1.30)	0.750	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.78 (1.21-2.64)	0.006	
	BCCA N=59	0.88 (0.57-1.36)	0.575	
	TCGA N=31	1.63 (1.02-2.60)	0.029	
Arm 8q	Neutral N=144	Reference		
	Amplified N=97	1.17 (0.84-1.64)	0.354	
	Deletion N=37	0.75 (0.46-1.23)	0.251	

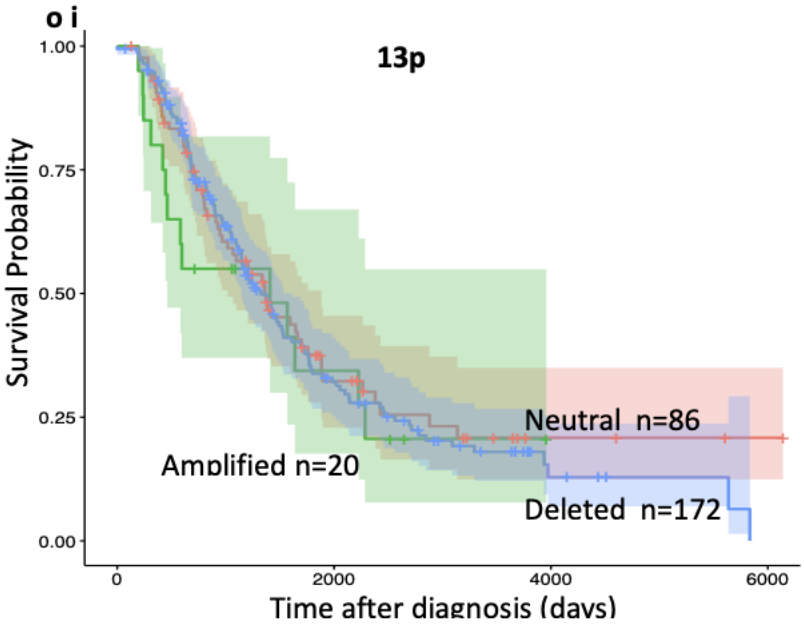
n i



n ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.071	
Stage	N=278	1.48 (1.13-1.95)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.69-1.27)	0.689	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.21-2.63)	0.003	
	BCCA N=59	0.89 (0.58-1.37)	0.597	
	TCGA N=31	1.67 (1.06-2.65)	0.028	
Arm 13p	Neutral N=57	Reference		
	Amplified N=29	1.12 (0.64-1.95)	0.702	
	Deletion N=192	1.02 (0.69-1.51)	0.91	

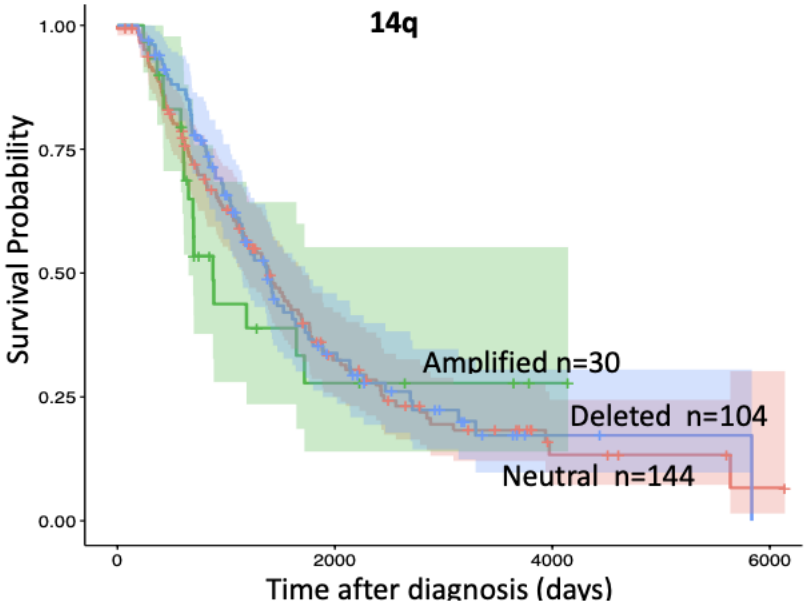
o i



o ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.082	
Stage	N=278	1.48 (1.13-1.96)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.33-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.92 (0.68-1.25)	0.596	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.81 (1.23-2.67)	0.003	
	BCCA N=59	0.89 (0.58-1.37)	0.609	
	TCGA N=31	1.68 (1.06-2.67)	0.028	
Arm 13p	Neutral N=86	Reference		
	Amplified N=20	1.12 (0.61-2.06)	0.719	
	Deletion N=172	1.15 (0.82-1.61)	0.417	

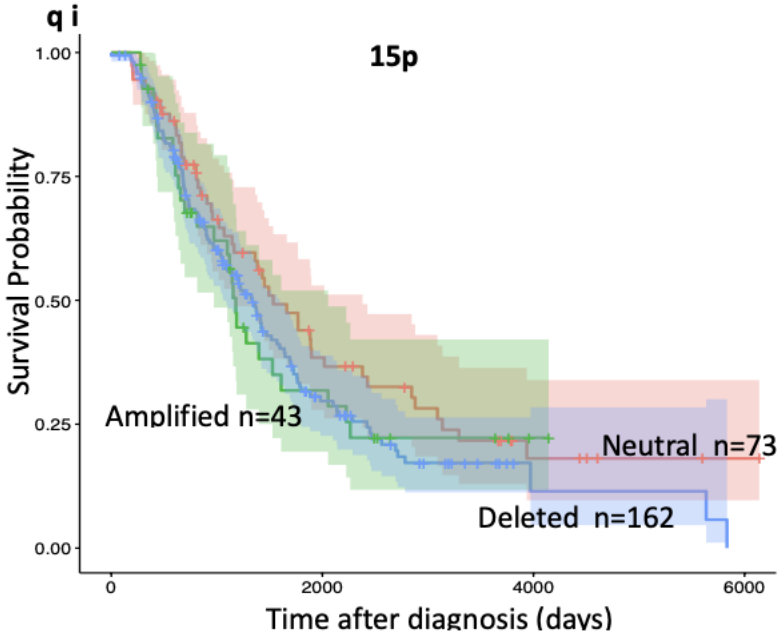
p i



p ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.034	
Stage	N=278	1.48 (1.12-1.95)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.45 (0.33-0.62)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.99 (0.58-1.36)	0.958	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.81 (1.23-2.66)	0.003	
	BCCA N=59	0.88 (0.58-1.36)	0.573	
	TCGA N=31	1.64 (1.04-2.59)	0.035	
Arm 14q	Neutral N=144	Reference		
	Amplified N=30	0.85 (0.50-1.45)	0.549	
	Deletion N=104	0.80 (0.57-1.11)	0.174	

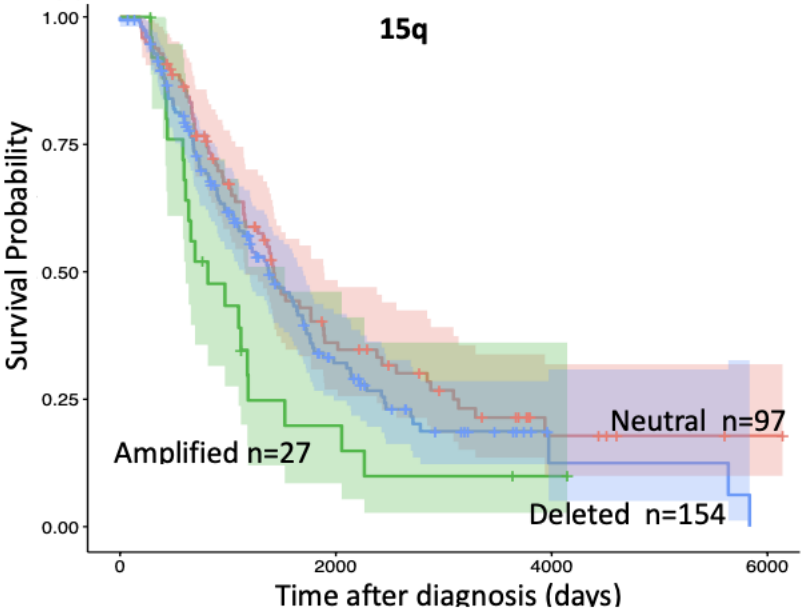
q i



q ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.066	
Stage	N=278	1.51 (1.15-1.99)	0.003	
HRD	Absent N=116	Reference		
	Present N=162	0.48 (0.35-0.66)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.88 (0.65-1.20)	0.415	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.88 (1.27-2.79)	0.002	
	BCCA N=59	0.91 (0.59-1.40)	0.671	
	TCGA N=31	1.72 (1.08-2.72)	0.021	
Arm 15p	Neutral N=73	Reference		
	Amplified N=43	1.33 (0.82-2.17)	0.250	
	Deletion N=162	1.36 (0.94-1.96)	0.106	

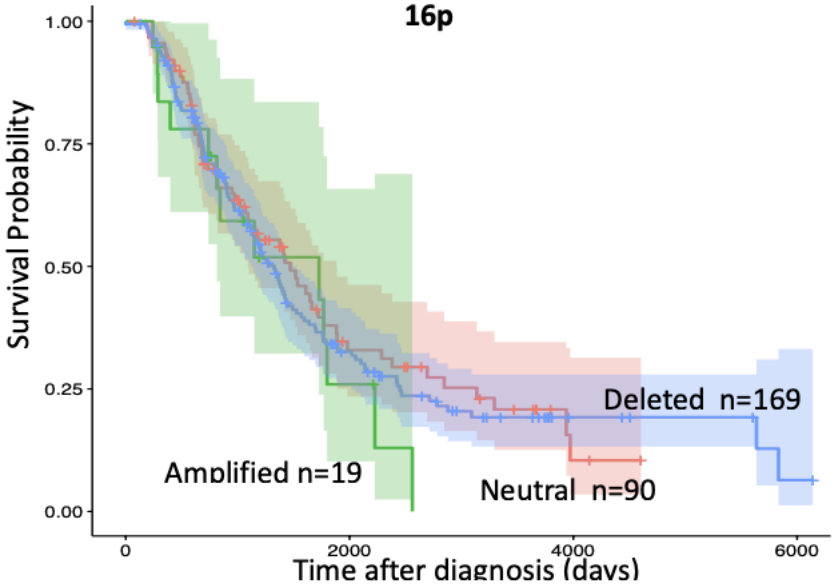
r i



r ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.065	
Stage	N=278	1.51 (1.15-1.98)	0.003	
HRD	Absent N=116	Reference		
	Present N=162	0.48 (0.35-0.65)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.91 (0.67-1.23)	0.538	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.883 (1.24-2.71)	0.003	
	BCCA N=59	0.85 (0.55-1.32)	0.475	
	TCGA N=31	1.68 (1.06-2.68)	0.028	
Arm 15q	Neutral N=97	Reference		
	Amplified N=27	2.02 (1.21-3.37)	0.007	
	Deletion N=154	1.21 (0.87-1.69)	0.260	

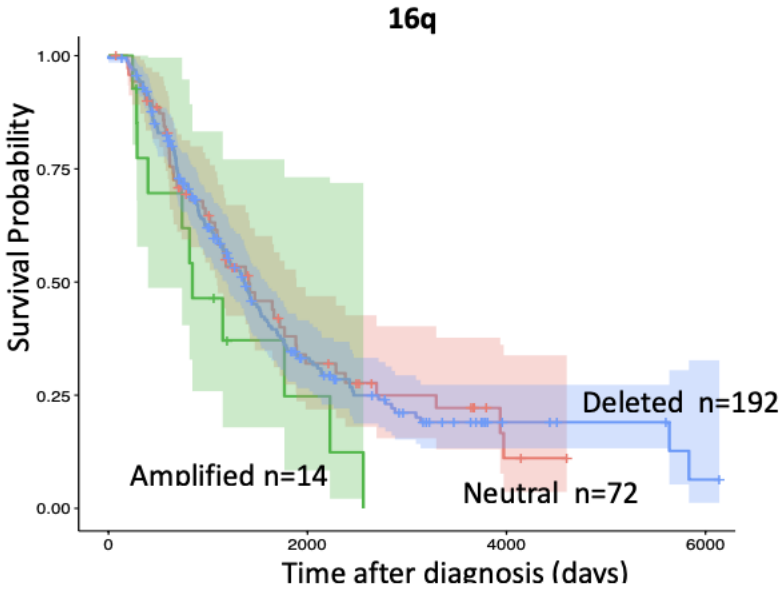
s i



s ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.062	
Stage	N=278	1.48 (1.13-1.95)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.95 (0.71-1.28)	0.732	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.21-2.65)	0.004	
	BCCA N=59	0.89 (0.58-1.37)	0.610	
	TCGA N=31	1.65 (1.04-2.63)	0.034	
Arm 16p	Neutral N=97	Reference		
	Amplified N=27	1.04 (0.56-1.92)	0.909	
	Deletion N=154	0.98 (0.71-1.35)	0.888	

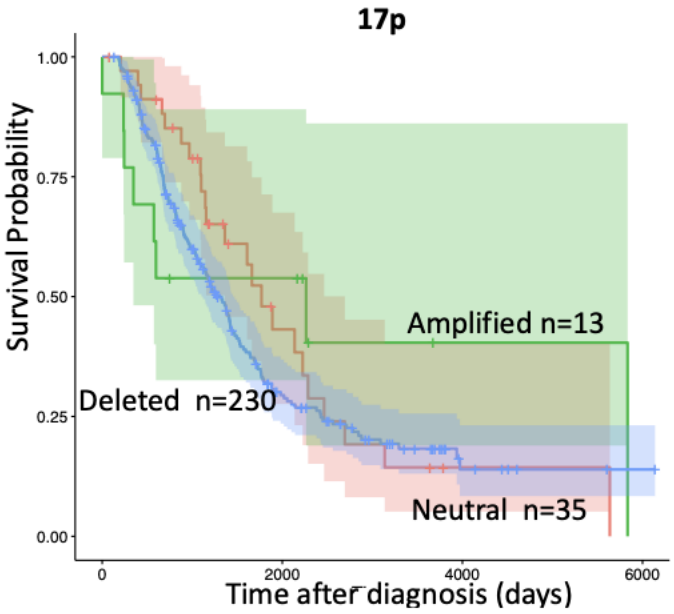
t i



t ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.046	
Stage	N=278	1.48 (1.13-1.95)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.95 (0.71-1.28)	0.752	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.21-2.65)	0.003	
	BCCA N=59	0.89 (0.58-1.37)	0.600	
	TCGA N=31	1.65 (1.04-2.63)	0.046	
Arm 16q	Neutral N=72	Reference		
	Amplified N=14	1.20 (0.61-2.37)	0.595	
	Deletion N=192	0.98 (0.62-1.23)	0.426	

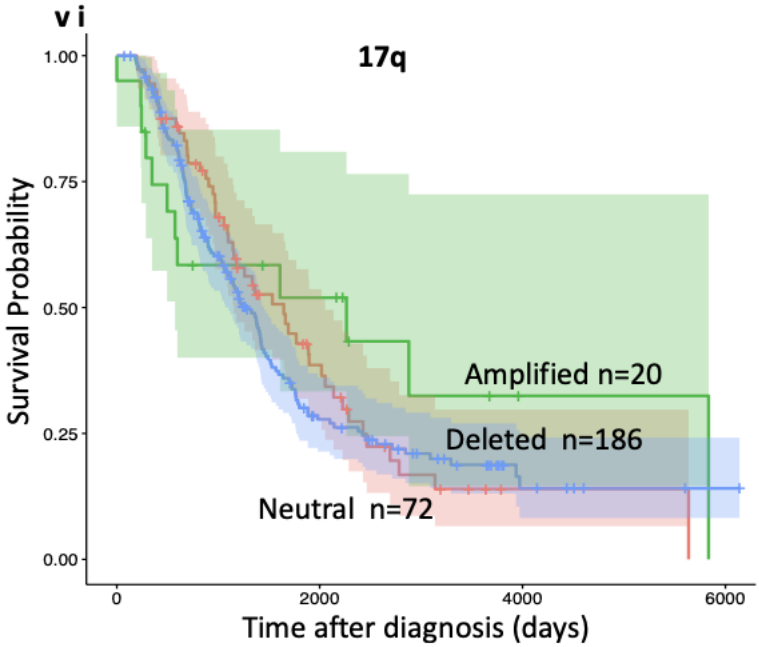
u i



u ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.042	
Stage	N=278	1.55 (1.17-2.05)	0.002	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.70-1.27)	0.705	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.72 (1.17-2.53)	0.006	
	BCCA N=59	0.88 (0.57-1.35)	0.555	
	TCGA N=31	1.62 (1.02-2.57)	0.039	
Arm 17p	Neutral N=35	Reference		
	Amplified N=13	1.31 (0.58-2.97)	0.521	
	Deletion N=230	1.41 (0.90-2.21)	0.133	

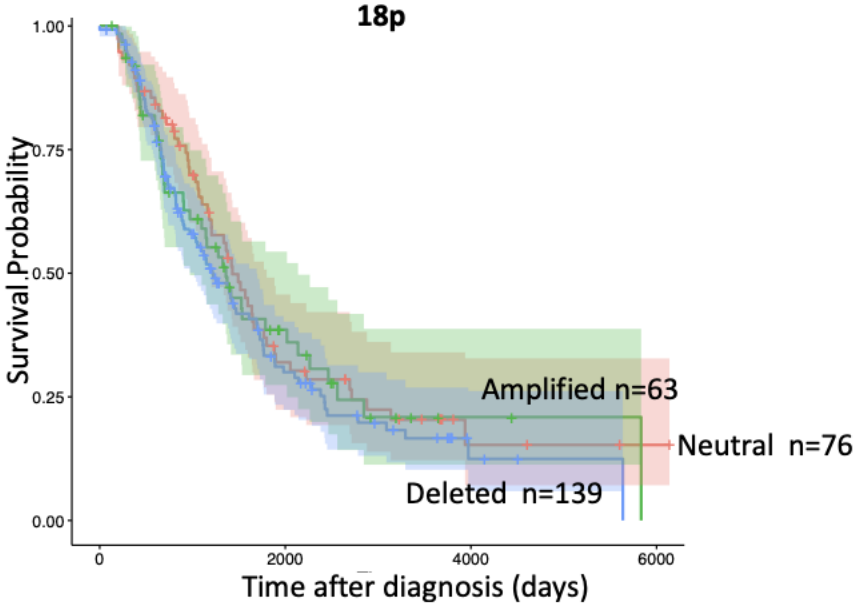
v i



v ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.051	
Stage	N=278	1.50 (1.14-1.98)	0.004	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.70-1.26)	0.668	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.78 (1.21-2.61)	0.004	
	BCCA N=59	0.90 (0.58-1.38)	0.622	
	TCGA N=31	1.64 (1.04-2.61)	0.035	
Arm 17q	Neutral N=72	Reference		
	Amplified N=20	1.01 (0.53-1.93)	0.964	
	Deletion N=186	1.16 (0.83-1.63)	0.376	

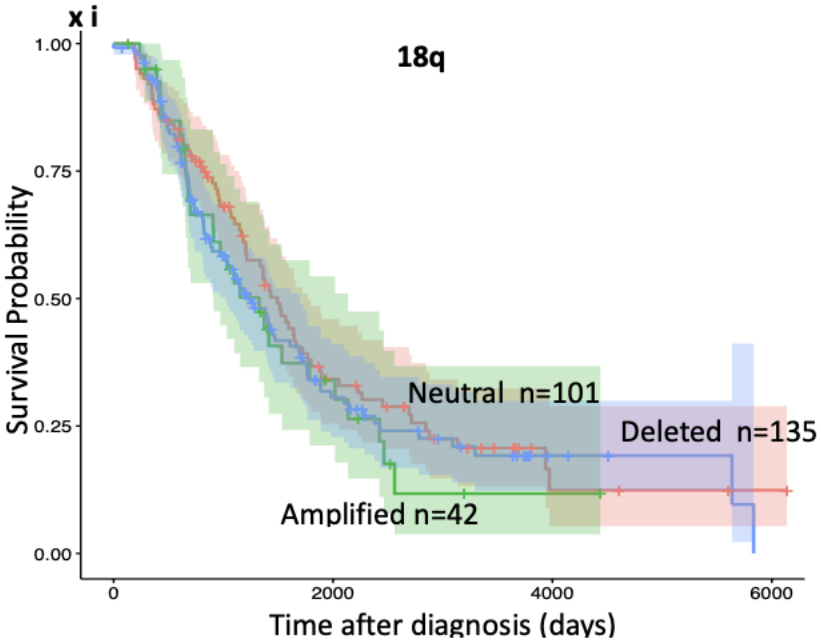
w i



w ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.072	
Stage	N=278	1.49 (1.13-1.97)	0.004	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.95 (0.70-1.27)	0.718	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.75 (1.19-2.58)	0.005	
	BCCA N=59	0.90 (0.58-1.38)	0.620	
	TCGA N=31	1.69 (1.07-2.68)	0.025	
Arm 18p	Neutral N=76	Reference		
	Amplified N=63	0.99 (0.65-1.93)	0.981	
	Deletion N=139	1.17 (0.82-1.65)	0.387	

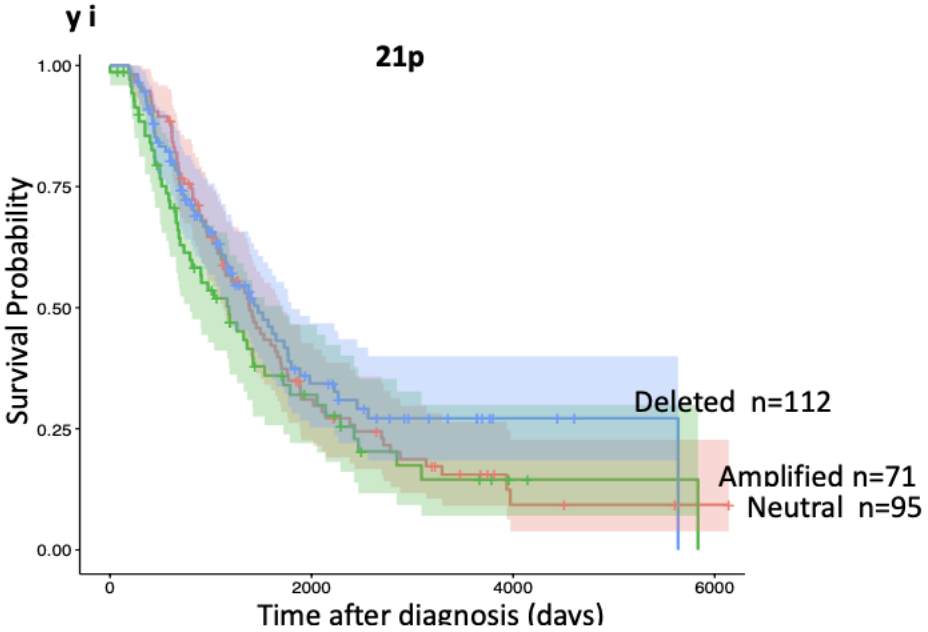
x i



x ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.083	
Stage	N=278	1.49 (1.13-1.96)	0.004	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.95 (0.71-1.28)	0.755	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.78 (1.21-2.61)	0.004	
	BCCA N=59	0.91 (0.59-1.40)	0.659	
	TCGA N=31	1.70 (1.07-2.69)	0.025	
Arm 18q	Neutral N=101	Reference		
	Amplified N=42	0.99 (0.63-1.55)	0.949	
	Deletion N=135	1.12 (0.81-1.55)	0.492	

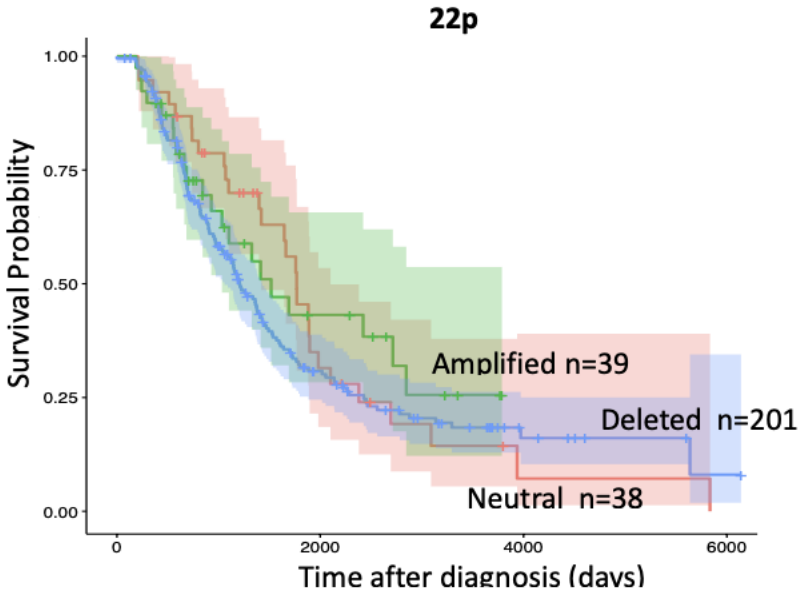
y i



y ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.056	
Stage	N=278	1.47 (1.12-1.94)	0.006	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.34-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.69-1.27)	0.680	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.72 (1.20-2.69)	0.005	
	BCCA N=59	0.90 (0.58-1.37)	0.613	
	TCGA N=31	1.71 (1.07-2.71)	0.024	
Arm 21p	Neutral N=95	Reference		
	Amplified N=71	1.24 (0.89-1.80)	0.249	
	Deletion N=112	0.97 (0.67-1.38)	0.847	

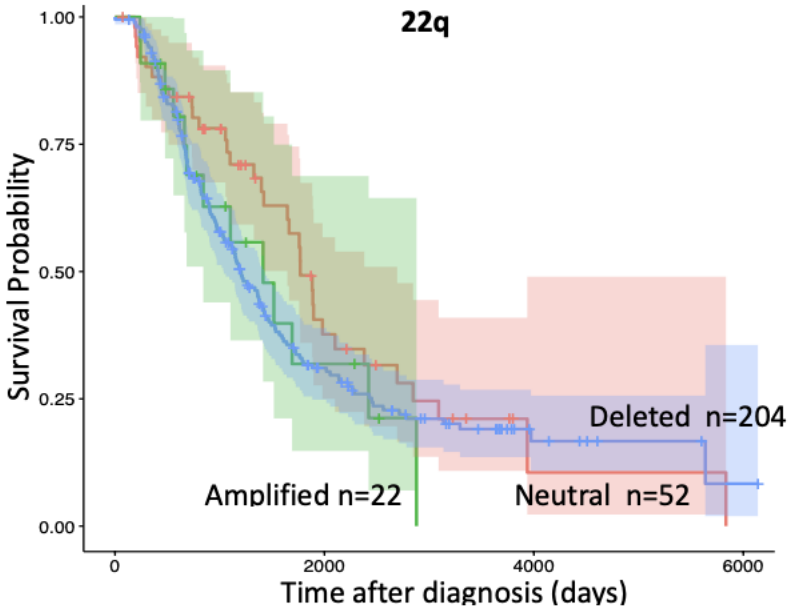
z i



z ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.063	
Stage	N=278	1.49 (1.13-1.97)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.71-1.29)	0.773	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.76 (1.18-2.60)	0.006	
	BCCA N=59	0.88 (0.57-1.36)	0.574	
	TCGA N=31	1.62 (1.02-2.60)	0.043	
Arm 22p	Neutral N=38	Reference		
	Amplified N=39	0.87 (0.48-1.56)	0.638	
	Deletion N=201	0.97 (0.62-1.46)	0.814	

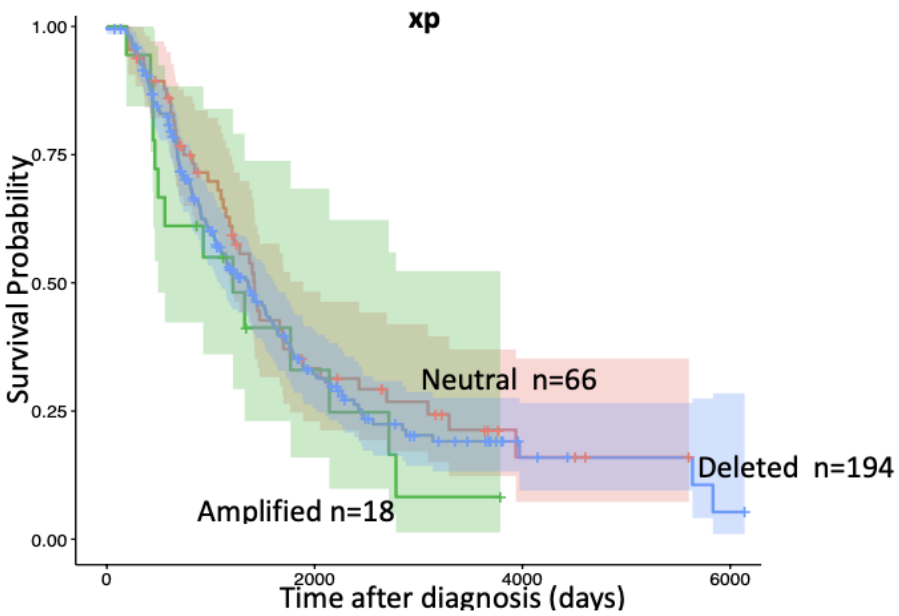
aa i



aa ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.02 (1.00-1.03)	0.061	
Stage	N=278	1.49 (1.13-1.96)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.96 (0.72-1.29)	0.795	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.81 (1.22-2.66)	0.003	
	BCCA N=59	0.90 (0.59-1.38)	0.636	
	TCGA N=31	1.70 (1.07-2.70)	0.024	
Arm 22q	Neutral N=52	Reference		
	Amplified N=22	1.35 (0.70-2.58)	0.371	
	Deletion N=204	1.03 (0.69-1.54)	0.896	

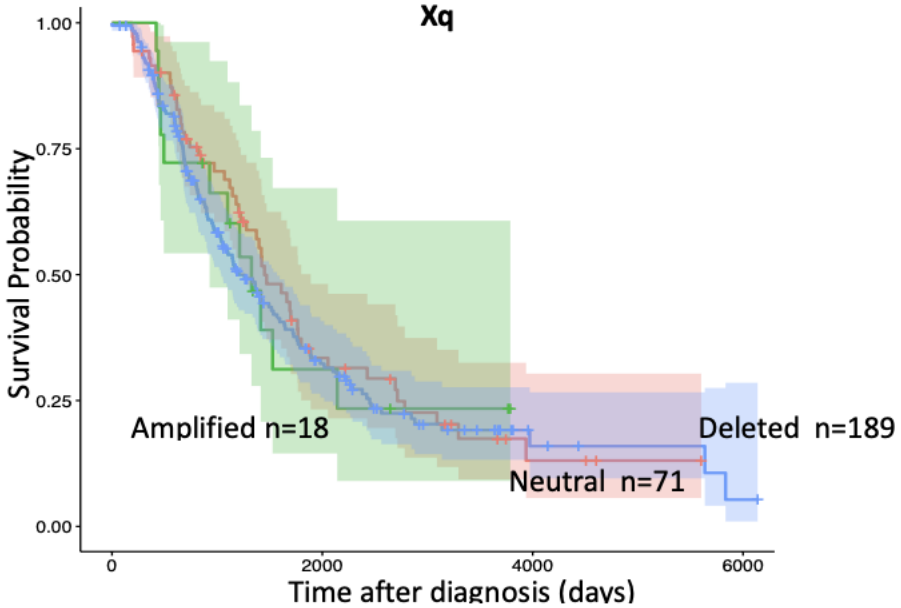
ab i



ab ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.061	
Stage	N=278	1.49 (1.13-1.92)	0.007	
HRD	Absent N=116	Reference		
	Present N=162	0.44 (0.32-0.61)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.92 (0.68-1.24)	0.583	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.80 (1.22-2.66)	0.003	
	BCCA N=59	0.90 (0.59-1.38)	0.616	
	TCGA N=31	1.66 (1.05-2.62)	0.031	
Arm xp	Neutral N=66	Reference		
	Amplified N=18	1.46 (0.77-2.75)	0.246	
	Deletion N=194	1.01 (0.71-1.44)	0.946	

ac i



ac ii

Condition		Hazard ratio	P value	Forest plot
Age	N=278	1.01 (1.00-1.03)	0.063	
Stage	N=278	1.48 (1.13-1.94)	0.005	
HRD	Absent N=116	Reference		
	Present N=162	0.46 (0.33-0.63)	0.001	
WGD	Absent N=139	Reference		
	Present N=139	0.94 (0.70-1.27)	0.688	
Cohort	SHGSOC N=108	Reference		
	AOCS N=80	1.79 (1.22-2.64)	0.003	
	BCCA N=59	0.91 (0.59-1.39)	0.646	
	TCGA N=31	1.68 (1.06-2.67)	0.027	
Arm Xq	Neutral N=71	Reference		
	Amplified N=18	1.22 (0.64-2.34)	0.543	
	Deletion N=189	1.05 (0.75-1.47)	0.792	

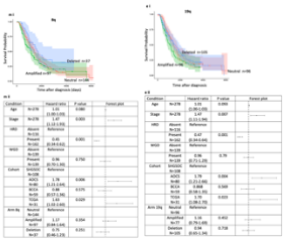


Figure 77 Impact of chromosome arm loss on survival

The frequency of chromosome arm amplification in samples with significant amplification shown in red **a**. The frequency of chromosome arm deletion in samples with significant deletion shown in blue **b**. Predictions of chromosome arm amplification or deletions where performed using GISTIC by Ailith Ewing (Mermel et al. 2011). For. each of the significantly amplified or deleted chromosome arms the impact on survival of amplification(green), deletion(blue) or neutral (red) arm changes are shown in a Kaplan-Meier survival curve for time after diagnosis in part **i** for panels **c** to **ac** A Cox proportion hazard model comparing impact on survival of amplification or deletion to neutral arm changes adjusting for age, stage at diagnosis, HRD status, and WGD a hazard ratio of 1 (no effect) is shown by a dashed line in part **ii** for panels **c** to **ac**. Multiple testing correction by Benjamini and Hochberg was performed on all p values.

Supplementary files

Sample level calls of cSV:
Description of all the cSV found per sample.

Region level calls of cSV:
Description of all the cSV found per sample with chromosome, start and end.

Severity of cSV:
Number of SV explained by each cSV type per sample

“Don’t be nervous, you don’t have to fight them.”- Mum