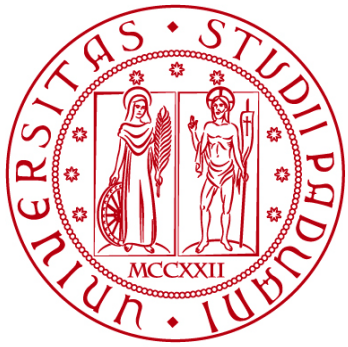
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology



TESI DI LAUREA

Somatic Copy Number Alteration detection and Copy Number signature analysis in High-Grade Serous Ovarian Cancer

Relatore: Prof.ssa Enrica Calura
Dipartimento di Biologia

Correlatore: Prof. Sampsa Hautaniemi
University of Helsinki, ONCOSYS Research group

Laureanda: Giulia Micoli

ANNO ACCADEMICO 2021/2022

Table of Contents

ABSTRACT

Somatic copy number alterations (sCNAs) are a type of genomic variation that affects the dosage of DNA sequences promoting tumorigenesis such as in High grade serous ovarian cancer. Their complexity prevents the unravelling of the mechanisms generating them and the molecular stratification of the patients. Here we propose the implementation of a highly sensitive pipeline for their detection based on structural variant calling and a signature analysis for the extraction of recurrent patterns. The precise calling allowed the recovery of small segments missed from the previous pipeline and in turn the clear distinction of patients with Homologous Recombination Deficiency (HRD), which resulted extremely segmented. Although the signature analysis, based on COSMIC copy number signatures, did not provide results consistent with the clinical data, the *de novo* signature extraction provided 15 new signatures able to proficiently explain the dataset. Two of them were positively associated with HRD, possibly representing a test for the identification of the HRD phenotype. Further insights on these signatures may provide the discovery of their etiology and give the possibility to shed light on association with single nucleotide variations.

CHAPTER 1 INTRODUCTION

THE HIGH GRADE SEROUS OVARIAN CANCER

Ovarian cancer is a type of tumor originating in the female reproductive organ and besides the common name, it denotes a multiplicity of distinct malignancies that share the anatomical site upon presentation.
Since it is the most lethal form of gynecological malignancies, it constitutes an impelling public health concern (8th cause of cancer-related death among women worldwide with a survival <30%).
One of the reasons for its high mortality is the late stage of the tumor at the moment of diagnosis and the lack of precise localization. Indeed, only 13% of diagnosed ovarian carcinomas are at stage I or II, the high majority have already metastasized.
There are currently no effective screening strategies for the early detection of ovarian cancer [1].

Ovarian Cancer Subdivision

The first classification dates back to 1930 WHO guidelines, which was based on histopathological differences. The 90% of ovarian tumors is estimated to derive from the transformation of epithelial cells and called Epithelial Ovarian Cancers (EOCs), as opposed to those originating from germ cells or sex-cord-stromal tissues. EOCs include four main histological subtypes determined according to morphology and tissue architecture: mucinous, serous, clear-cell and endometrioid.
Serous ovarian cancer has been further classified by the introduction of grading systems for a more accurate prognosis. The 2-tier grading system is the most famous and based on two histological assessments, nuclear atypia (how much tumor cells differ from normal tissue) and mitotic rate. The classes defined are:

1. Low-grade serous carcinoma (LGSC): low-grade nuclei (look almost like normal cells) with infrequent mitotic figures (tend to grow slowly). It evolves from adenofibromas or borderline tumors. The progression to an invasive type is slow and stepwise. It is indolent and has a better outcome than high-grade.
2. High-grade serous carcinoma (HGSC): high-grade nuclei and numerous mitotic figures. It is an aggressive malignant neoplasm without obvious precursor lesions.

More recently, EOC has been characterized according to genetic and molecular features and the new classification has been recognized by WHO in 2014. The first division is in two broad categories, Type 1 and Type 2.
Type 1 neoplasms are characterized by a stepwise development mainly from pre-malignant or borderline lesions as other epithelial cancers and typically present a large cystic neoplasm. From the genetic point of view, they are stable, P53 wild type and expose alterations in RAS-MAPK and PI3K-AKT signaling pathways. Type I includes low-grade serous, clear-cell and mucinous subtypes.
Type II malignancies develop more rapidly, are usually widely disseminated at the time of presentation and are more aggressive. They share P53 mutations and genomic instability because of DNA repair

pathway defects. The dominant subtype of this category is the high-grade serous ovarian cancer, which accounts for 70-80% of deaths from all forms of ovarian cancer [2].

This project focuses in particular on the study of HGSC, therefore further descriptions will be all related to this subtype.

Origin and development of HGSC

Ovarian cancer is a heterogeneous group of neoplastic diseases. According to the latest revisited and revised model of ovarian carcinogenesis, serous tubal intraepithelial carcinoma (STIC) is thought to be a potential precursor of HGSC. STIC is most probably the earliest morphologically recognizable precursor lesion of many (not all) pelvic HGSC. The exact mechanism, how the STIC develops into invasive pelvic serous carcinoma is not well understood [1].
HGSC does not require blood or lymphatic systems to disseminate and metastasize since it extends to adjacent organs within the peritoneal cavity or through detachment of cells from primary tumor. The exfoliated cells, singles or in clusters, are suspended in the peritoneal fluid, spread, catch on distant organs or tissues and grow.
The peritoneum is a membrane lining the abdominal cavity, covering the inside wall of the cavity and every organ contained in it. The peritoneal cavity in normal situation contains 50-75 ml of serous peritoneal fluid. Its functions are mainly support of the viscera, insulation, lubrification, blood, lymph and nerve supply, and immunity (barrier to pathogens). To fulfill its functions, it is developed into a highly folded, complex structure.
The folds of the peritoneum divide the abdominal cavity in several compartments, among which there is the omentum. The omenta are folds enclosing nerves, blood vessels, lymph channels and fatty and connective tissues.
The favorite metastatic site of HGSC is the omentum, most commonly without affection of the underlined organs and with colonization only of the mesothelial cell layer. Patients usually are diagnosed with a late-stage disease, which also presents ascites (abnormal accumulation of fluid within the abdomen) and in the fluid the presence of spheroids or aggregates have been proposed to represent a unit of metastatic spread [2].

Genomic context of HGSC

The Cancer Genome Atlas (TCGA) is a project begun in 2006 from the joint collaboration of the National Cancer Institute (NCI) and the National Human Research Institute. It is a cancer genomics program whose aim is the molecular characterization of primary cancer and matched normal samples spanning 33 cancer types.
TCGA studies allowed to shed light on the genomic and transcriptomic landscape of many tumors, including EOC. The vast majority of EOC TCGA samples are HGSC [3]. Starting from these data, it has been increasingly clear that:

1. the most prevalent somatic mutation across patients affects *TP53* gene with a missense mutation producing a truncated protein, indicating that most likely it is required for the initiation of the disease and enables the copy number pattern;

2. other important mutations include *BRCA1/2* (role in DNA repair, recombination, and transcription), *CSMD3* (regulation of dendrite development), *NFI* (Nuclear Factor), *CDK12* (transcription regulator of DNA repair genes);
3. genomic instability characterizes this cancer type which is reflected in high number of copy number alterations, the result of which is the amplification of deletion of many genes; such as *CCNE1* (cyclin, regulatory subunit of CDK2 for G1/S transition), *MYC* (nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis, and cellular transformation) and *MECOM* (transcriptional regulator and oncoprotein that may be involved in hematopoiesis, apoptosis, development, and cell differentiation and proliferation);
4. considering pathway analysis for HGSC pathogenesis, it is possible to notice that the homologous recombination repair (HRR) pathway is highly defective in most patients, involving mutations in *BRCA1*/2 and other genes (*RAD15*, *PTEN*, *RAD51C*, *ATM*, *ATR*). It represents a central high-fidelity DNA damage-repair system responsible for reparation of DNA double-strand breaks and interstrand crosslinks in a slow, specific, complex, and accurate fashion. The mutation of this pathway allows the distinction of two classes of patients: homologous recombination deficient (HRD) and homologous recombination proficient (HRP). In HRD patients, the repair of DSB is given by non-homologous end-joining, single-strand annealing or microhomology-mediated end joining pathways, which are not as effective and exact as HRR [4].
   Additional mutated pathways comprise Notch, *PI3K*, *RAS-MEK* and *FOXM1* signaling.

HGSC Treatment

The paradigm for newly diagnosed HGSC is that they are treated by primary surgical cytoreduction followed by platinum-based chemotherapy or in case of a too large dissemination of the tumor or tumor mass located in a wrong site the patient undergoes a primary neoadjuvant chemotherapy (NACT) before the cytoreduction and additional chemotherapy [2].
The primary surgical cytoreduction or "debulking" aims to remove all tumor masses in the peritoneal cavity. The more advanced the disease stage, the more complex the operation and the less likelihood of success. When NACT is required, it consists of three cycles of carboplatin and paclitaxel followed by interval, surgical cytoreduction and additional chemotherapy. It is used especially when the patients are too ill to undergo surgery or when the cancer burden is too extensive.
Recommendations for the use of adjuvant chemotherapy using platinum-based chemotherapy for patients with early-stage ovarian cancer depend on the stage, grade and histology (patients with grade I are not treated with chemotherapy post-surgery, but higher grades do). Its effectiveness has been improved by combining cisplatin and carboplatin with taxanes, anti-angiogenic agents (bevacizumab) and other drugs.
Cisplatin and carboplatin belong to the class of alkylating agents, which cause DNA damage adding bulky alkyl to guanine nucleotide bases and therefore inhibiting proper DNA synthesis. Taxanes were first isolated from

the cortex of *Taxus brevifolia*, inhibit tubulin depolymerization and cause dysregulation of the cell cycle with mitotic failure [1].

Relapse and Platinum resistance

Although there is a good response for platinum-based chemotherapy, more than 80% of patients will relapse (platinum resistant patients). For these patients, alternative or second-line drug combinations are utilized and then followed-up for 2-4 months with physical examination or radiographic imaging [2].

The time between last platinum chemotherapy and recurrence, known as the platinum-free interval (PFI), represents one of the most important prognostic factors. Traditionally, patients recurring six months or more after last platinum are labeled platinum sensitive and most of them respond positively with a rate between 30-90% to further platinum-based chemotherapy. Patients recurring <6 months from last platinum are considered platinum resistant and typically have low response rates to additional chemotherapy. Nevertheless, in general patients retreated with platinum-based chemotherapy manifest response rates of almost 50% but the efficacy decreases, together with life expectancy [1].

Platinum resistance is one of the main causes of the high mortality of this disease, but the mechanisms are still poorly understood. Some hypotheses involve drug influx and efflux pathways, intracellular redox balance and drug modification, tumor microenvironment, DNA damage repair machinery, epigenetic mechanisms. For example, mutations in BRCA1/2, genes of the homologous recombination (HR) process, have been shown to be associated with sensitivity to platinum therapy and reversion mutations which restore their function confer resistance to platinum-based drugs and PARP inhibitors [5].

Germline mutations in BRCA1/2 seem to favor a longer survival rate and high responsiveness to platinum-based therapies. In these patients targeted therapies are used and one of the most common approaches relates to PARP inhibitors. When HR deficiency is present, the repair is over-reliant on the poly (ADP-Ribose) polymerase (PARP) mediated base excision repair (BER). Drugs targeting PARP are very effective in this case because the cell fails completely to repair DNA breaks[2].

Whitin platinum resistant patients, a part will positively respond to further platinum-based chemotherapy but the prediction of who will benefit from it is still an unsolved problem. Moreover, the establishment of different treatment lines and the choice of which of them to use is different from individual to individual and is dependent on the history of prior treatment, residual toxicities and the availability, cost and convenience of treatments [6].

THE COPY NUMBER VARIATIONS

Structural Variants

One of the most important mutational processes in cancer are the structural variations (SVs). SVs are defined as sequence variants >50bp in size but they can vary a lot in type and size (from 50 to Mbp) they comprise a lot of subclasses consisting of unbalanced Copy Number Variants (CNVs), which include deletions, duplications, and insertions of genetic material, balanced rearrangements such as inversions, interchromosomal and intrachromosomal translocation, but also include mobile elements insertion, multi allelic CNVs, segmental duplication, complex rearrangements.
One SV is identified as a junction between two breakpoints in the genome and usually there is also a change in copy number across the breakpoint if only one side of the break is rescued by a structural variant, otherwise it results in a balanced structural variant [7].

Among all classes of structural variants, the current project is focused on Copy Number Variations since they largely affect HGSC tumor genomes and are caused by genomic instability.

sCNAs: what are them?

Copy Number Variations are regions of the genome that vary in integer copy number, meaning that there is a different number of copies of that region from *2n* (human normal ploidy) [8].
Nevertheless, some lexical formalities have been proposed that distinguish them depending on the size of the event [9]:

- Aneuploidy is defined as a CNV event affecting entire chromosome arms, whole chromosomes or even whole genome;
- CNVs describe all sub-arm gains or losses larger than 10kb;
- Indels describe all other CNVs (shorter than 10kb).

CN events can be further classified into germline and somatic events. Moderate and physiological germinal CNVs can be detected in all healthy individuals. They are particularly important in the evolutionary context, in which they can generate biodiversity over long time scales, driving rapid adaptations in response to stress and change in the environment. Although somatic CNV, also called somatic Copy Number Alterations (sCNAs), are the CNVs occurring in somatic cells that can occur because of not-detected defects in DNA replication of recombination. Determining the CNV functional consequences in both germline and somatic cells is challenging because they result in alleles of large effect that impact more genes and regulatory regions at the same time [8].

sCNA in cancer

In cancer, high levels of sCNA occur whenever the transformed cell accumulates mutations in genes of DNA replication and repair pathways in HR pathways. Somatic CNA can be both the cause or the consequence of cell transformation [8]. sCNAs represent the type of mutation that affects the largest part of the genome in cancers.

Two main challenges in understanding sCNAs in cancer are recognized:

1. Distinction between driver events (responsible for oncogenesis) and passenger mutations (acquired secondarily). Driver mutations confer clonal growth advantage and are positively selected. One possibility is assuming that passenger mutations are randomly distributed but it is not always correct. A cluster of somatic mutations may also be attributable to a local increase in mutation rate. Moreover, most copy number alterations involve loss or gain of broad chromosomal regions. For example, CN loss targeting a tumor suppressor gene can also involve multiple neighboring genes not involved in cancer development. The loss of the neighbors can render cancer cells vulnerable to further suppression or inhibition of those genes. To shed light on this point, the mechanistic point of view must be considered.
2. Identification of target genes (oncogenes or tumor suppressors) of driver sCNAs and determination of their functional role. The context is examined in this case because positive correlations of sCNAs with other genetic events may indicate functional synergies, while anticorrelations functional redundancies [10].

It is clear that sCNAs are pervasive across human cancers and characterize certain tumor types. They can carry prognostic information and they can reflect the level and type of genomic instability. What is still not known is how they form in the first place and how or if they evolve during tumor progression.
Another challenging question is whether sCNAs affect the spatial arrangement of the genome in the nucleus changing the gene expression and cell fitness. Genomic rearrangements are likely to cause a repositioning of the genes and regulatory regions at a local level or affecting the overall architecture. Understanding this point would allow us to gain a deeper knowledge of mechanisms conferring higher fitness to tumor cells.
Lastly, sCNAs must be studied because they could represent a novel therapeutic target [10].

sCNA Detection

The research in SVs field in cancer encounters difficulties due to biological factors among which are tumor heterogeneity, purity, and polyploidy.
Tumor biological samples are never composed of only tumor cells but there is always contamination of normal cells. Purity represents the percentage of cancer cells over the total and is an important parameter to understand if a variant is germline or somatic.
Moreover, a tumor mass is composed of different types of cancer cells that derive from the acquisition of a progressive level of mutation because of genomic instability. This leads to high heterogeneity between cancer cells forming subclones. Subclonality is a problem because the presence of variants that are present only in a small number of cells in the samples decreases the power of detection.
Polyploidy also represents an obstacle because the overall tumor ploidy influences the calculation of alleles values [7].

High-throughput sequencing (HTS) techniques are nowadays applied for copy number variations detection. Many of the prevalent tools and

algorithms are SV callers which use short reads to infer the presence of SVs compared with a reference genome.
Although short-read approaches are highly effective at resolving single nucleotide variants (SNVs), SV detection is unable to completely overcome the read sequence and insert sizes of standard short-read HTS.
The methods are several but in general all consist in the identification of mapping discordance between the HTS read and the expectations given by reference genome. The features considered are:

- Read depth (RD): changes are associated with sCNAs causing genomic rearrangements. Genomic fusion partners cannot be identified, and breakpoint position is not precise.
- Discordant aligned read Pairs (DP): pair read aligned with unexpected orientation or separation, or to different chromosomes.
- Split Reads (SR): the sequenced reads span the breakpoints. Methods using them find breakpoints by identifying split alignments in which part of the read aligns to either side of a genomic rearrangement in three possible ways: direct split read mapping, realignment of Soft-Clipped (SC) bases (unaligned bases in partially mapped reads), split alignment of the unmapped read in One-Ended Anchored (OEA) read pairs
- Local assembly: assembly of reads obtained from clusters of SCs or OEAs pairs to form break-end contigs (extend and span out the breakpoint).

Not all these features are used by all SV callers but the integration of more of them can significantly increase the power and sensibility of detection [11].

Focusing on CN, the calling algorithms rely mainly on the segmentation of the genome, that is the partition into regions with a distinct copy number profile. The approaches are generally three, among which the last one is the most used: Hidden Markov Model (HMM), Circular binary segmentation (CBS) and rank segmentation.
After the segmentation two parameters important for sCNAs evaluation are calculated: purity and ploidy. Purity represents the percentage of the tumor cells over the total amount (each sample is prone to contamination of normal cells in the tissue) while ploidy is the number of chromosomes occurring in the nucleus of a cell (it can be altered in tumor condition). They are fundamental for the calculation and correction of B Allele Frequency (BAF) and Log R ratio (logR).
BAF and logR metrics allow calculation of the final CN for both alleles and the visualization of the CN pattern in the genome.
BAF is the measure of the minor allele and its % at each position in the normal condition (diploidy) has three possibilities: 0 (AA homozygous), 100 (BB homozygous) and 50 (AB heterozygous). If there is a copy number variation (gain or loss) or the sample is not pure, the values change. It is a measure of the allelic contrast and is also useful for the visualization of Loss of Heterozygosity (LOH) cases.
Log R ratio instead, is a metric that represents the signal intensity for CNV analysis. This parameter originates from a polar coordinate transformation used in microarrays of two-channel intensity data. This transformation generates a normalized intensity value called R and an allelic intensity ratio.

The intensity comparison is done by looking at the observed normalized intensity of the subject sample (Rsubject) compared to the expected intensity (Rexpected) and computing the base two logarithm of their ratio [12].
In the copy number case, for each position, the expected number of calls is calculated and compared to the observed one. The result is the logarithm of the ratio between the two amounts. If it is >0 the number of calls observed is higher than the expected one, meaning a gain in the number of copies, while if <0 the opposite (loss of copies). This measurement must be taken into account to confirm CNVs, especially LOH cases (B allele alteration can be due to other reasons) [13].

MUTATIONAL SIGNATURES

Mutations in cancer genomes are caused by mutational processes (exogenous or endogenous origin) that impact on cell lineages between zygote and the formed cancer cell. Every mutational process involves the modification of specific pathways in the cell and therefore, when it activates the repair mechanisms, these generates characteristic patterns of mutations in the genome [14]. From the point of view of genomic architecture, locations of somatic mutations are modified by replication timing, transcriptional activity, eu- and hetero-chromatin presence, histone modifications, transcription factor binding site presence; while from sequence perspective mutational processes differ in biophysical and biochemical characteristics, resulting in a specific preference for the sequence context of the somatic mutations. The combination of these two aspects affects the accumulation of mutations leading to a characteristic pattern [15]. These patterns are called mutational signatures and can involve Single Base Substitutions (SBSs), Double Base Substitutions (DBSs), small insertions and deletions (IDs), genomic rearrangements and copy number changes (CN).
Identification of these signatures and their cause is an arduous task and requires mathematical modeling because each cancer genome might have been generated from a combination of mutational processes that have to be discerned [14].
The main repository of officially accepted mutational signatures is the COSMIC (Catalog Of Somatic Mutations In Cancer) database, divided into several projects among which there is the signatures part (https://cancer.sanger.ac.uk/cosmic) [16]. Provided signatures aren't definitive but represent a reference set of high confidence. They have been identified from the analysis of PCAWG dataset and curation of specific papers and consist in four variant classes for which mutational profile, proposed etiology and tissue distribution are provided: SBS, DBS, ID (indels) and CN.
The clinical utility of signatures derives from the possibility to understand the mutational processes shaping and modifying cancer genomes and therefore identifying a number of environmental mutagens. As a consequence, these findings allowed the development of strategies for decreasing the exposure to such mutagens.
They also helped to understand the impact of therapies which lead to selection of resistant clones that can cause secondary or recurrent cancers

and, in the end, allow us to understand tumor mass evolution processes [15].

Copy Number signatures in HGSC

Copy Number signatures have been proposed only recently because mutational processes that drive copy number changes are not readily identifiable from genome-sequencing data. One of the first studies showing the utility of CN signatures focuses on HGSC. Shallow Whole Genome Sequencing (WGS) data are used in a mixture modeling approach to separate copy number features distributions and then non-negative matrix factorization (NMF), being able to extract seven signatures. Mutational processes have been associated with each signature and it has been shown that they are also able to predict the overall survival [17].
In June 2022 two pan-cancer studies focusing on copy number signatures were published. The first one focuses on chromosomal instability in human cancer, looking at characteristic genomic patterns and using data from TCGA obtained with the Affymetrix SNP Array 6.0 platform (microarray technology). 17 signatures have been identified and associated with probable causes, but they were also able to predict drug response and platinum sensitivity. One problem with this study is clearly the use of an old technology and the fact that the method doesn't show concordance in the identification of the same signatures in WGS data [18].
The second study produces the actual 21 copy number signatures in COSMIC, identifying some associated with LOH, HRD, chromothripsis or driver genes. WGS and Whole Exome Sequencing data have been used and decomposed with the same method used for substitutional and ID signatures. The approach enables the identification of both shared patterns of copy number across all examined samples and the quantification of the number of segments attributed to each copy number signature in each sample [19].
Given the data type used and the reliability of the paper, COSMIC signatures have been chosen for the quantification of segmentation data and successive analysis.

GOAL OF THE PROJECT

The study of Copy number alteration in cancer through NGS technique is extremely important because it supports precision oncology clinical research. Indeed, sCNAs can be used for the diagnosis, prognosis and treatment of HGSOV.

To accomplish this task, improving the detection of these variants is necessary. One of the objectives is the implementation of a pipeline for the identification of the somatic copy-number variations in ovarian tumor samples maximizing the sensibility and specificity of the tools used.

The following step aims to find common patterns in the distribution of sCNAs in tumor samples, that is a signature quantification analysis. This passage is important because it helps in finding shared alterations that can describe subgroups of patients (HRD/HRP, high LOH ..) and can be targeted for a possible treatment.

CHAPTER 2 METHODS

Cohort description

The cohort consists of 233 patients suspected of ovarian cancer, of which 195 confirmed in HGSC mainly at the III-IV stage (95%). All patients participating in the study provided written informed consent. The study and the use of all clinical materials have been approved by the Ethics Committee of the Hospital District of Southwest Finland (ETMK) under decision number EMTK: 145/1801/2015.
The clinical specimens used in the study represent several understudied aspects of HGSC that are poorly represented in existing cohorts of clinical specimens, such as TCGA. The Cancer Genome Atlas (TCGA) is a cancer genomics project which molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Cancers for study have been chosen based on specific criteria that include poor prognosis, overall public health impact, availability of samples meeting standards for patient consent, primary, untreated tumor with a source of matched normal tissue or blood sample, frozen, sufficiently sized, resection samples, and samples composed of at least 80% tumor nuclei (threshold later lowered to 60% with improved sequencing technology and computational methods).
Contrary to TCGA data, all samples were collected from intra-abdominal, peritoneal, and omental metastases, thus representing cancer cell populations with proven metastatic potential. The cohort also included low purity tumors that may represent a distinct, poor prognosis phenotype of HGSC, which are missing from most genomic analyses.

Purity in samples has been determined in most part of the sets only computationally and not assessed initially by a pathologist. This choice comes because sequenced samples are fresh-frozen, and the matching formalin fixed with paraffin embedding (FFPE) have not been produced. The sequencing of FFPE samples is not optimal because of short inserts and artificial base alterations which lead to overestimation of certain variables [20]. The last set has also been provided of hematoxylin and eosin images of the samples, permitting to estimate the purity.
Nevertheless, the low purity samples are retained and considered for the analysis because most of them represent the interval phase after the treatment which implies that also in poor responders there is a decrease of the purity of the sample. Moreover, purity also depends on the site of origin of the samples: in ovary, obtaining high purity samples is relatively easy, while in metastatic sites like peritoneum the samples are more often characterized by a low purity.

The age at diagnosis ranges from 38 to 88 years old with a mean of 68. Half of the patients have been treated with primary debulking surgery (PDS) and platinum-taxane chemotherapy and the other half with NACT therapy. Moreover, half of the patients have been associated with SBS3 mutational signature, associated with HRD, and half without it (HRP).
The medium platinum-free interval (PFI) is 362 days.

Data pre-processing

Fresh tumor samples are collected from surgeries (primary tumor, metastases, ascites) and plasma samples during 1st line treatment and disease relapse in Turku University Central Hospital (Tyks). The samples are first processed at the University of Turku, where DNA, RNA, circulating tumor DNA (ctDNA) extraction and quality control are performed. Sequencing is entrusted to the Novogene sequencing center which executes a further step of quality control, library preparation and sequencing. Samples are sent and divided in batches called "sets" and the sequencing characteristics can vary.
WGS data sequenced before mostly 30x coverage (BGISEQ and HiSeq), now primarily 50x (Novaseq) with Illumina HiSeq X Ten, BGISEQ-500 or MGISEQ-2000, or Illumina NovaSeq 6000. The data are paired-end reads at 150 (HiSeq, NovaSeq, BGISEQ) or 100 bp (set 6 BGISEQ) in FASTQ format.

Table 1: Summary of sequencing features used for the samples.

Platform	Number samples	Set	SAM platform	Instrument
HiSeq	204	3,4,5,6	Illumina	HiSeq X Ten
BGISEQ	399	6,7,8,9	DNBseq	BGISEQ-500, MGISEQ-500
NovaSeq	540	9,10,11,12	Illumina	NovaSeq 6000

Bulk WGS data are preprocessed by quality control and trimming (QCFasta, using FastQC, trimmomatic 0.3), alignment to the reference genome (GRCh3.8.d1.vd1) with BWA-MEM version 0-7.12-r1039, duplicate marking (Picard MarkDuplicate version 2.6), base quality score recalibration (GATK BaseRecalibrator version 3.7) and cross-sample contamination estimation (GATK version 4.1.9.0). There isn't a proper normalization for correcting the possible batch effect, arising because of the different technologies used, but a manual control is done, and some statistics are calculated and controlled for consistency for each technology used (e.g. number of mutations and indels).
Downstream analysis performed involve germline and somatic short variant detection (GATK version 4.1.9.0), copy-number alterations and tumor purity estimation (GATK version 4.1.4.1 with ASCAT algorithm) and attribution of mutational signatures (COSMIC v3.1).

CURRENT PIPELINE

GATK: segmentation

The currently used workflow for the sCNAs analysis consists in three steps: (i) segmentation of WGS data using GATK, (ii) ploidy and purity estimation using ASCAT algorithm and (iii) visualization of the results with Genome Spy.

Genome Analysis ToolKit (GATK) is a collection of command-line tools for analyzing high-throughput sequencing data with a primary focus on variant discovery. The tools can be used individually or chained together into complete workflows [18].. Main steps:

1. Data preprocessing: each individual read-pair is mapped to the reference genome producing a SAM/BAM format sorted by coordinate, read pairs that are likely to have originated from duplicates of the same original DNA fragments are marked and samples are corrected for patterns of systematic errors in the base quality scores.
2. Variant discovery: starting from the BAM file variants are called per-sample. Depending on the target type of variant different tools can be used. The genomic VCF intermediate files are combined per sample into a multi-sample genomic VCF file and after a final step of refinement the VCF result is produced.
   For sCNAs the key is looking at the coverage profile to determine whether the variation occurred or not. The main step is the coverage normalization and then the segmentation, where the boundaries of the events (segments having the same uniform copy number) are determined. In somatic variants the task is more difficult because the size of the events is very small.
3. Variant filtering: since variant callers are sensitive, identifying potential false positives and applying filters to remove those less likely to be real variants is necessary. The steps include variant quality score recalibration, hard filtering on quality criteria and using annotation features [21].

For a sCNA analysis, before collecting coverage counts, the resolution of the analysis is defined with a genomic intervals list. Preparing a genomic intervals list is necessary whether an analysis is on targeted exome data or whole genome data. In the case of whole genome data, the reference genome is divided into equally sized intervals or bins and raw integer counts data are then collected.
BAFs are collected using all filtered biallelic germline SNPs with heterozygous calls from each patient. Read-count collection used one kilobase intervals. Both read and allelic count collection excluded regions listed in the ENCODE blacklist [22] and internal DECIDER blacklist. The DECIDER blacklist includes regions that have abs(logR) > 0.2 in at least three of the available normal samples. The 136 regions in the DECIDER blacklist represent poorly aligned regions and population-level copy-number variance.
A Panel of Normals (PoN) is also required. The normal samples in a PoN should match the sequencing approach of the case sample under scrutiny.

The PoN stores information that, when applied, will standardize case sample counts to PoN median counts and remove systematic noise in the case sample. The read counts are standardized and denoised against the PoN producing the standardized copy ratios.
The workflow creates a somatic Panel of Normals from the WGS normal blood samples of patients given a list of BAMs. Only samples with less than 5% contamination are included. Interval-specific and platform-specific Panels of normal are created first and then concatenated in the final one.
In segmentation, contiguous copy ratios are grouped together into segments. The tool performs segmentation for both copy ratios and for allelic copy ratios, given allelic counts. Counts of the reference allele and counts of the dominant alternate allele are tabulated for each site in a given genomic intervals list and the collection for the case and the matched-control alignments is done independently with the same interval. In the somatic case, the matched-control is the germline normal sample and the case is the tumor sample from the same individual.
Copy and allelic ratios that are contiguous on the same segment are then grouped together.
The Gaussian-kernel binary-segmentation algorithm enables efficient segmentation of dense data, like that of whole genome sequencing. The algorithm performs segmentation for both copy ratios and for allelic copy ratios jointly when given both data types together. Systematic calling of copy-neutral, amplified and deleted segments is finally performed [21].

ASCAT: purity and ploidy

The ploidy and purity estimation are performed with ASCAT (Allele Specific Copy number Analysis of Tumors). The idea is the automation of the discovery of ploidy and contamination of samples for sCNA analysis. ASCAT procedure involves both segmentation, variant calling and estimation of purity and ploidy. Nevertheless, since the segmentation is performed with GATK, ASCAT algorithm has been modified and adapted to perform only the calculation of the parameters [23].
Purity and ploidy (ρ and ψt) are looked so that the allele-specific copy number estimates are as close as possible to a non-negative whole number for germline heterozygous single nucleotide polymorphisms (SNPs). Allele-specific copy number profiles are calculated for a grid of purity and ploidy values, for each one the distance to a positive integer number is calculated and local minima are established. Goodness of fit is calculated as a linear rescaling of the total distance to nonnegative whole numbers to a percentage and used to choose the most likely solution.
At the end, with the solution of purity and ploidy provided by the calculations, it is possible to calculate the corrected BAF and LogR and the estimates for the allele-specific copy numbers [21].

Defects of GATK-ASCAT pipeline

There are two main reasons for the implementation of a new workflow for sCNA calling.
First, GATK estimates the segmentation singularly for every sample, even if more samples belong to the same patient, thus they have the same genetic background and clonal origin. When the samples of a single patient are

compared the call of the breakpoint is often not precise and can slightly vary between one sample and another even if they are most likely the same breakpoint. The consequence is a generation of noise in the visualization and estimation of the boundaries of the segments.
Moreover, because of tumor heterogeneity and contamination from normal cells, it is possible to find in the same sample only a few cells that contain a specific variant. The inspection of the BAM file allows its recognition, but GATK tool is not able to detect it because of the low number of supporting reads. All these problems could be solved by applying a joint calling for the breakpoints determined by SV detection, read depth and BAF. Joint calling consists in considering all reads coming from the samples of the same patient together to call the structural breakpoints (as they come from a single sample). It ensures that a common variant near the single-sample threshold of detection will be reliably reported as a shared variant. Joint calling allows for sensitivity detection of variants that are present subclonally (or at low coverage) that would not be detected if called individually. Joint calling has higher coverage of shared variants thus resulting in more reliable assembly of that variant.

IMPLEMENTED PIPELINE

The attempt to improve the GATK-ASCAT pipeline has been accomplished by the use of tools implemented by Hartwig Medical Foundation (HMF), a project born in the Netherlands in 2015 with the aim of performing DNA analysis of cancer patients. They were able to generate a database storing the clinical and genetic information of thousands of patients with metastatic cancer. They also developed a state-of-the-art IT pipeline for the bioinformatic data analysis together with various software tools to call, analyze and annotate the WGS data. All code used (including all relevant documentation) can be accessed on Github: https://github.com/hartwigmedical/.

HMF tools range from identification of viral integration to detection of structural variations, to single-nucleotide variations and more. For sCNA analysis the tools used (Fig.1) and their main functions are:

- GRIDSS: call of SVs between tumor/reference pairs (not developed by HMF);
- GRIPSS: provide a somatic filter for the extraction of only somatic variants, also removing low quality calls;
- Amber: determines the BAF of heterozygous germline variants in the tumor samples;
- Cobalt: extracts read depth ratio;
- Purple: combines output from all other tools to produce the somatic output

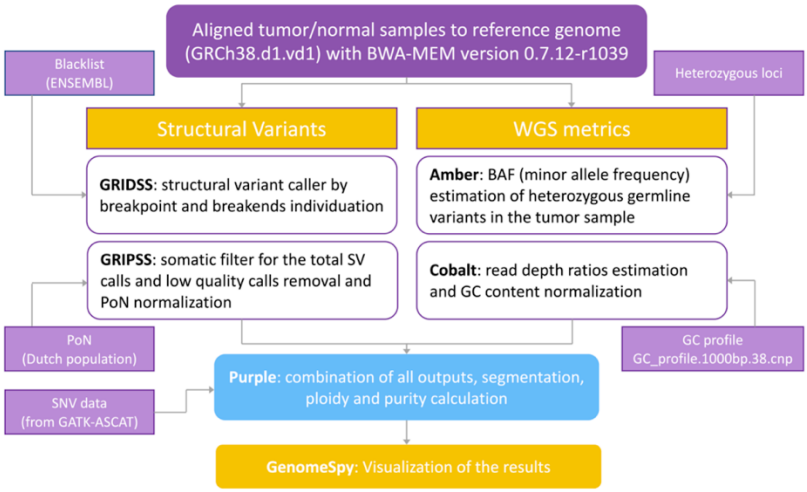


Fig 1: Schematic representation of the HMF pipeline. The input of all tools is represented by the BAM file. Processes are parallelized and divided in two branches, one identifying structural variants and one calculating the WGS metrices. SVs are called by GRIDSS using also the ENSEMBL blacklist and its output is passed to GRIPSS which, using a PON, carries out the somatic filtering. At the same time Amber takes as input the heterozygous loci and calculates the BAF while Cobalt uses the GC profile of each sample for normalization and calculation of the read depth. The outputs of GRIPSS, Amber and Cobalt are used by Purple, together with SNV data for the segmentation and the estimation of purity and ploidy. Segmentation is then visualized in GenomeSpy.

GRIDSS

GRIDSS (Genomic Rearrangement IDentification Software Suite) is a module software suite containing tools useful for the detection of genomic rearrangements. It provides a multithreaded variant calling from a combination of assembly, split read and read pair support. It maximizes the sensitivity and prioritizes calls into high or low confidence, thereby maintaining specificity in the high-confidence call set.
It takes a three-step approach:

1. Filtering out reads that align properly: all reads providing any evidence for underlying genomic rearrangements are extracted
2. Assembly of the extracted reads (genome-wide break-end assembly): each contig formed corresponds to a break-end and only after assembly, the underlying breakpoint and the partner break-end are identified. It assembles all SC, SR, DP, OEA and indel-containing reads. Breakpoints are identified by realignment of break-end contigs.
3. Variant calling: a probabilistic model that combines break-end contigs from each side with SR and DP evidence to score and call variants

To perform the genome-wide break-end assembly, a positional de Bruijn graph data structure has been extended. Positional de Bruijn graphs were first developed for small indel and base calling error correction of *de novo* assembly contigs. The main advantage of these graphs is the inclusion of positional information in the nodes of classic graphs.
Joint calling is suggested to be always used for related samples (tumor-normal matched samples). Indeed, GRIDSS performs joint assembly using all related samples as they are one for the calling of SVs, but then reports the extracted variants per-sample.

The output consists mainly in a VCF file with break-ends reported, accompanied by a quality score file. Each call is a breakpoint consisting of two break-ends, one from location A to location B, and a reciprocal record from location B back to A [24].

HMF tools

GRIDSS Post Somatic Software (GRIPSS) is a tool which applies a set of filtering and post processing steps on GRIDSS paired tumor-normal output to produce a high-confidence set of somatic SVs for a tumor sample. It processes GRIDSS output and produces a somatic VCF.
The filters applied are hard filters (exclusion of no mates, set a tumor minimum quality, set a minimum support for a variant). GRIPSS realigns the variant to the earliest possible base in the uncertainty window which is the most likely base for the soft clipping. Soft filters are then applied [25].

Amber is designed to generate a tumor BAF file for use in PURPLE from a provided VCF of likely heterozygous SNP sites (WGS heterozygous sites determined using GATK). When using paired reference/tumor BAMs, AMBER confirms these sites as heterozygous in the reference sample BAM, then calculates the allelic frequency of corresponding sites in the tumor BAM. To do that, it performs a segmentation step using the Piecewise Constant Fitting algorithm (PCF) [25].

Cobalt (Count bam lines) tool determines the read depth ratios of the supplied tumor and reference genomes. It starts with the raw read counts per 1000 base window for both normal and tumor samples by counting the number of alignments starting in the respective BAM files with a mapping quality score of at least 10. It then applies a GC normalization to calculate the read ratios.
The reference sample ratios have a further 'diploid' normalization applied to them to remove megabase scale GC biases. This normalization assumes that the median ratio of each 10Mb window (minimum 1Mb readable) should be diploid for autosomes and haploid for male sex chromosomes.
It also performs a segmentation step using the PCF algorithm exactly like Amber [25].

Purple (Purity and ploidy estimator) combines B allele frequency (BAF) from Amber, read depth ratios from Cobalt, somatic variants and structural variants to estimate the purity and CN profile of a tumor sample (final segmentation). To run, it also requires the same GC profile as used in Cobalt and a reference genome.
Providing the GRIDSS SVs set as input allows the exact base resolution of copy number changes and a high set of somatic SNV and indel calls can also improve the accuracy.

It functions in 12 steps, among which the most important are:

1. Segmentation: it combines segments of read ratios from Cobalt and BAF points from Amber with SV breakpoints using specific rules. Every SV starts a new segment and ratio or BAF breaks are included only if they are at least one mappable window away from an existing segment. Once segments have been identified, in each one median tumor BAF, median read ratio and number of BAF points are recorded. A reference CN status is also determined as diploid, heterozygous deletion, homozygous deletion, amplification or noise.
2. Sample purity and ploidy: it considers a matrix of all possible combinations and scores each one on a segment-by-segment basis. The aim is finding the most parsimonious solution for the fit. The penalties used include sub-clonality, solutions which deviate from diploid heterozygous CN, solutions with implausible somatic SNV copy number, weight by count of BAF observations and giving more weight on segments with higher observed BAF. A fit score is calculated and the solutions within 10% or 0.0005 of the best are retained as candidates.
3. Copy number smoothing: initial segmentation is very sensitive and the read depth from WGS is noisy. Therefore, many adjacent segments will have similar BAF and CN profiles but they unlikely represent a real change. For these reasons, a smoothing algorithm is applied to decide for segment merging or not.
4. Allele specific copy number: it is possible that some regions lack BAF points and result in an unknown allele-specific CN (since BAF coverage is limited). BAF and allele CN are inferred from neighboring regions with known values and from observed copy number changes to the unknown regions.

Other steps are also present such as structural variant recovery, identification of germline gene deletions, QC status for the tumor.

The final output consists in a number of tab-separated files containing values for the summary purity, best fit purities sorted by score, copy number profile of all contiguous segments of the tumor sample, significant amplification and deletions that occur in the HMF gene panel…
It also provides a list of VCF files if structural or somatic VCF files have been supplied to PURPLE. Corresponding VCF are written to the output directory enriched with purity information [25].

Comparison: GenomeSpy visualization and ploidy differences

The first comparison implied the estimation of the ploidies coming from the different tools. By a ggplot (R package) scatterplot it has been possible to visualize all data points with respect to the ploidies and see the ones which differ most.

The second comparison involved the visualization of the segmentation by using GenomeSpy. GenomeSpy is a visualization toolkit for genomic and other data developed in the Systems Oncology group at the University of Helsinki. It has a Vega-lite inspired visualization grammar and high-performance graphics rendering.

It is split into five packages but the most important are the core and the app. The core library provides the visualization grammar and the JavaScript programming interface. The app extends the visualization grammar with support for faceting many patient samples. It provides a user interface for interactive analysis of the samples, which can be filtered, sorted, and grouped. It supports data in CSV, TSV, JSON and FASTA format.
For the purpose of the project, it has been used for visualizing and comparing the segmentation of all samples thus it is possible to catch the differences in the results of the two pipelines. Moreover, it has also been exploited to create a more detailed view of the segmentation of a single sample by showing all data points as a function of the LogR and the segments from both the pipelines.

COPY NUMBER SIGNATURE ANALYSIS

COSMIC copy number signatures have been extracted in three steps:

- Copy number profile summarization
- Copy number signature identification
- Assignment of the signature to single samples

Copy Number profiles summarization

Segments are classified hierarchically according to three features:

- Heterozygosity state: if the copy number of the alleles is A>0 (major allele) and B>0 (minor allele) the segments is heterozygous, if A>0 and B=0 the segment has LOH, if A=B=0 the segment is characterized by homozygous deletion;
- Sum of major and minor allele (TCN): TCN=0 homozygous deletion, TCN=1 deletion leading to LOH, TCN=2 wild type and copy neutral LOH, TCN=3/4 minor gain, TCN=5-8 moderate gain, TCN≥9 high-level amplification;
- Segment size: 0–100kb, 100kb–1Mb, 1Mb–10Mb, 10Mb–40Mb and >40Mb.

Using this categorization, copy-number profiles are described as counts of 48 combined copy number features defined by heterozygosity, copy number and size[19].
The copy number profiles of the dataset are thus summarized as a non-negative matrix with S x 48 dimensions (where S is the number of samples) using SigProfilerMatrixGenerator.
SigProfileMatrixGenerator is a computational package written in Python with an R wrapper package that allows the efficient exploration and visualization of mutational patterns (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator). It is able to read somatic mutational data in most commonly used data formats such as Variant Calling Format (VCF) and Comma-Separated Values (CSV). It was first created for the classification and analysis of SBS, DBS and ID, but then adapted to create a matrix also for CNs. Indeed, given as input the signatures table downloaded from COSMIC, the mutational profiles and the reference genome, it transforms the mutational catalogs of the samples into the mutational matrices called $\boldsymbol{M}$, and outputs them as text files [26].
To avoid wrongly estimating the signatures, samples with a purity lower than 0.2 have been filtered out because the segmentation is not reliable.

Copy number signature identification

Given the mutational matrix, the global reference set of copy number signatures was used to assign an activity for each signature to each sample of the dataset. For this purpose, three different methods have been used:

- Decomposition module of SigProfilerExtractor;
- Poisson model: selection of the signatures based on BIC (Bayesian Information Criterion) using Poisson likelihood;
- Non-Poisson model: selection based on the cosine similarity.

The assumption of all methods is that the mutational spectrum of a tumor can be represented as a linear combination of signatures.

SigProfilerExtractor is a method based on nonnegative matrix factorization (NMF) (https://github.com/AlexandrovLab/SigProfilerExtractor). NMF is the approximate representation of a nonnegative matrix $\boldsymbol{M}$, in this case the observed mutational profiles of a set of samples, as the product of two usually smaller nonnegative matrices, $\boldsymbol{W}$ and $\boldsymbol{H}$, which are the signatures and the attributions respectively. NMF is calculated from 256 to 1024 times with different random initial conditions because the profiles of the signatures can vary substantially depending on the input samples and because with multiple similar signatures operating, there are multiple possible reconstructions [14].
SigProfilerExtractor implementation can be separated into seven modules packaged together into a single tool:

- Module 1: input processing
- Module 2: resampling and normalization of the mutational matrix
- Module 3: multiple NMF replicates
- Module 4: hierarchical clustering to perform model selection
- Module 5: decomposition of *de novo* signatures into COSMIC signatures
- Module 6: calculation of activities in individual samples
- Module 7: outputting and plotting [27]

For signatures attribution and quantification, only module 6 is used with COSMIC derived signatures. The combination of Module 5 and 6 has been recently detached, forming a separate tool called SigProfilerAssignment (https://github.com/AlexandrovLab/SigProfilerAssignment). It attributes a known set of mutational signatures to an individual sample or multiple samples, decomposes de novo signatures to COSMIC database and attributes COSMIC database or a custom signature database to given samples. It identifies the activity of each signature in the sample and assigns the probability for each signature to cause a specific mutation type in the sample [27].

Dataset with short segments

The COSMIC signatures are estimated from the classification of the segments as previously described. As far as segment length concerns, the category with the shortest segments range collects segments from 0 to 100kb because the categories have been created as respect to ASCAT results [19]. As explained, the new implemented pipeline allows to increase the power of detection and the sensibility of the variant calling, meaning that also very short segments are identified [24]. Given this, it is possible that the

categories established for the copy number signatures do not correctly represent these segmentation data. To verify it, a new segmentation dataset has been created in which segments shorter than 10kb are removed and the neighboring segments are merged. The new logR and BAF are calculated by weighted average of the merged segments.
The signature analysis is repeated also in this dataset.

Comparison and visualization

The exposures calculated from the different methods are then compared and quantified for how much they explain the provided data. To accomplish this task, the cosine similarity is calculated for each sample comparing the vector of the exposures and the vector derived from the multiplication of the signature matrix and the feature matrix (***M***). The average of the cosine similarity of all samples is then calculated. The result is a number between 0 and 1 that represents how much the exposures represent initial data.
This metrics is calculated for all the three methods used and also for the dataset with the short segments.

Visualization of the signatures attributions is then provided with bar plots (ggplot R package) that for each sample show the present signatures and its amount and with a GenomeSpy visualization.

De novo extraction

To properly fit data coming from the used cohort, extraction of new signatures has been performed. For signatures building, the initial dataset of samples has been filtered, selecting for each patient one sample per phase; if more samples are present within the same phase, the one with the highest purity is chosen. This selection is necessary to eliminate the bias caused by the different number of samples per patient, but maintaining the biological variation introduced by the treatment phases.
Before proceeding with the extraction, a new category has been added to the previously described used to extract COSMIC signatures. The first class of segment size has been split in two distinct classes: 0-10kb and 10kb-100kb. The final number of categories at this point is 58. The purpose of this operation is to better fit Purple segmentation.
Copy number profiles are summarized in a matrix with the same technique described but using the modified version of SigProfilerMatrixGenerator.
Signatures are extracted using all seven modules of SigProfilerExtractor (v.1.0.17). It first performs independent Poisson resampling of the input matrix for each replicate to ensure that fluctuations of the matrix do not impact the stability of the factorization results. A step of normalization is also applied to overcome potential skewing of the factorization from any hypermutator. The second phase involves the matrix factorization which factorizes the matrix given as input with different ranks, searching for an optimal solution between k=1 and k=40 signatures. For each value of k it runs 500 independent NMFs of the normalized Poisson resampled input matrix by minimizing an objective function based on the Kullback-Leibler divergence measure. The result consists in k sets of matrices, each one containing 500 different matrices ***H***, each reflecting the patterns of de novo

mutational signatures and 500 matrices *W*, each reflecting the activities of the de novo signatures.
At this point, hierarchical clustering is applied to the 500 factorizations to identify the stability of decomposition (whether solutions from different initial conditions converge to similar signatures). It selects the centroids of stable clusters as optimal solutions, making these solutions resistant to fluctuations in the input data [27].

CHAPTER 3 RESULTS

SEGMENTATION

Pipeline output

Purple, from the implemented pipeline, produces for each sample eight files. The purity file, called TUMOR.purple.purity.tsv, contains a single row with a summary of the purity fit: purity, score of fit, the average ploidy of the tumor sample after correction for purity and other features. The purity range file (TUMOR.purple.purity.range.tsv) instead summarizes the best fit per purity sorted by score, meaning that it reports all grid possibilities analyzed by the algorithm.
From these files the pipeline builds a sunrise plot for each patient, which shows the range of scores of all examined solutions for purity and ploidy. Ploidy is represented as a function of the purity and for each combination, the score is represented in a color scale. The best solution corresponds to the point with the lowest score and is indicated by the cross of the dashed lines (Fig2).
The possible patterns that can be present are typically three:

- The purity is very low, meaning that the sample is composed almost totally of normal tissue. As a consequence, the sunrise plot is characterized by many stripes with low score values in correspondence of 2, 4, 6 values of ploidy (Fig2a). Normal samples and samples with a whole genome duplication (WGD) have a ploidy corresponding to one of these values and it is not possible to identify which is the correct one because they all fit equally. This is the case also of cancer cells that have not copy number aberrations but present only SNVs.
- The sample presents multiple solutions for the ploidies and thus there are many blue spots in the plot (Fig2b). The solution with the lowest score is chosen by the algorithm but the other options are still probable.
- The plot shows only one minimum, corresponding to the optimal solution and the best score (Fig2c). In this case, the estimation of the solution is easier.

Together with these files, Purple also produces a copy number file for each sample named TUMOR.purple.cnv.somatic.tsv. It contains the copy number profile of all contiguous segments of the tumor sample, thus the chromosome, start and end of the segment, BAF, GC content, copy number of major and minor alleles adjusted for purity and the type of structural variant support for the copy number breakpoint at start and end region.
Other important output files consist in the structural variant VCF and somatic variant VCF which contain all entries from the input structural variant VCF (GRIPSS output) and somatic VCF (somatic SNV data) enriched with some fields for the purity adjusted allele frequency, copy number and change in copy number and others.

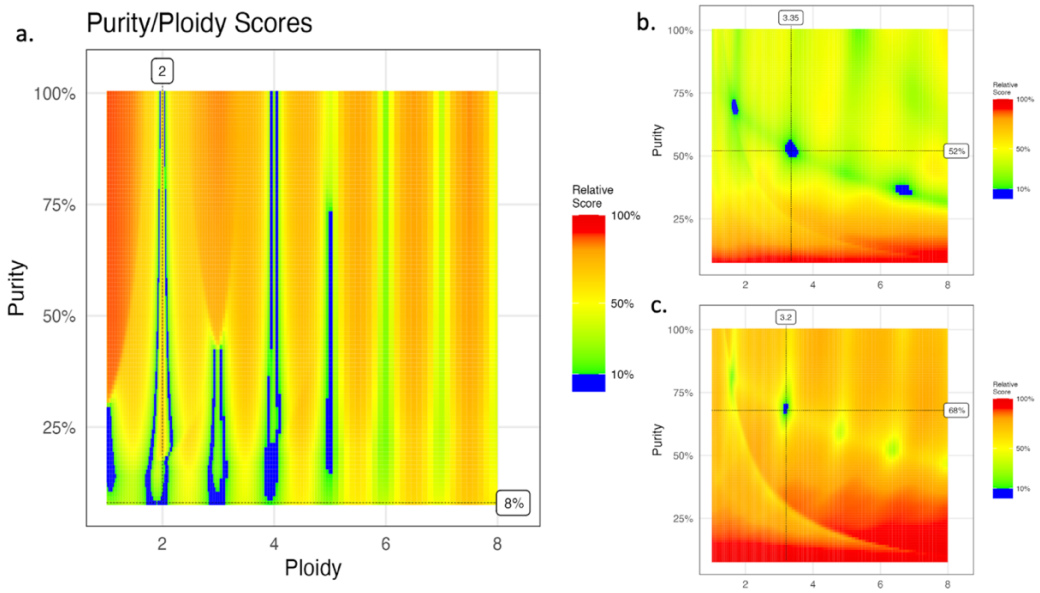


Fig 2: Purple sunrise plots. Sunrise plots represent the scores for each ploidy-purity combination of the grid used by the algorithm. Axes represent the ploidy, expressed as integer numbers, and purity, expressed as percentage. The color scale indicates the value of the score transformed as a percentage. The lower the score, the better the estimation. Values under 10% are all represented with the same color. The vertical and horizontal dotted lines indicate the position and the valued for the best estimate. A) Low purity samples display this pattern, which mimics the situation of a normal sample with vertical stripes in correspondence of ploidies equal to 2 or multiples. B) Multiple-ploidy sample. More than one solution result to be an optimal estimate, thus there are more blue areas. C) One-ploidy sample. There is only one plausible solution indicated by the blue area.

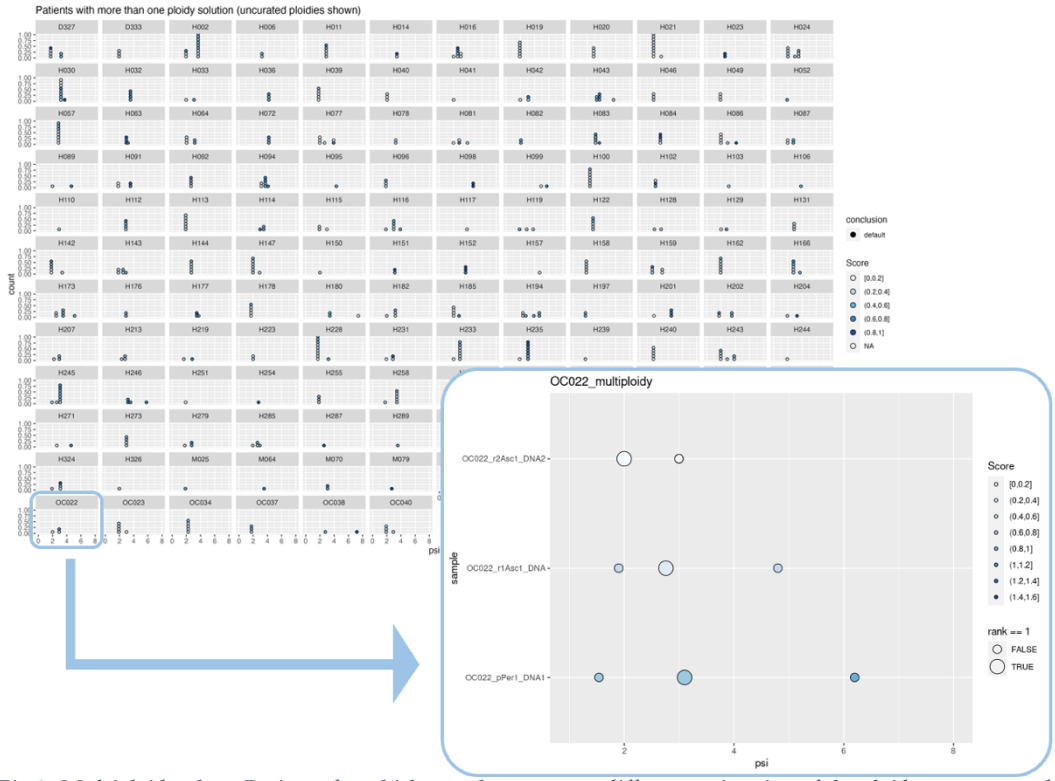


Fig 3: Multiploidy plots. Patients for which samples present a different estimation of the ploidy are extracted and the best solutions for the ploidy are reported. In the main plot all samples are represented, and each dot represents a solution, the color scale indicates the score (the lighter, the better). In the zoom the plot for the single patient is shown. For each sample (y axis) the best ploidies are shown (x axis), the color represents the score, and the larger dot is the best solution chosen by the algorithm.

An important analysis of the patients with multiple possible ploidies has also been conducted. Patients which have samples presenting different estimated ploidies are extracted and the most plausible ploidy solutions for each sample have been represented. Functions to build the plots and find the minimas of the scores have been taken from the ASCAT algorithm. One plot for each multiploidy patient is provided (Fig3 zoom) together with a general one (Fig3).

The plot for the single patient shows all its samples (y axes) and the minima of the ploidy found by Purple algorithm (x axes, psi). The chosen solutions are highlighted by a bigger dot, while the color represents the range of the score (the lower the better, meaning the lighter the better).
The plot with all patients represents a summary of all chosen ploidies. The color scale for the score is maintained and the dots represent the first solution of all samples in each patient.
These graphical representations consent to determine if the estimation of the ploidy is plausible or not according to the values of all samples in the same patients, the tissue site of the sample and its phase. For example, in Fig4a the sample in the first row has an estimated ploidy of about 2 but the score is not very reliable. The samples coming from omentum and ovary (pOme1 and pOvaL1) show a ploidy closer to 3 and, considering that these are the most likely sites where the tumor originated, it is more probable that the ploidy of the peritoneum (preferred site of metastasis) is the same. This is also supported by the fact that the range of the score is similar between the two points. It is also possible to check it from the related sunrise (Fig4b) which shows in the two red circles that both the points are quite reliable (it isn't possible to clearly see it because with a score lower than 10% the color is blue with no gradient). The manually curated ploidy is shown in green in Fig4a.
Overall, analyzing all created multiploidy and sunrise plots, there are just few samples that necessitated a manual curation and estimates seem plausible.

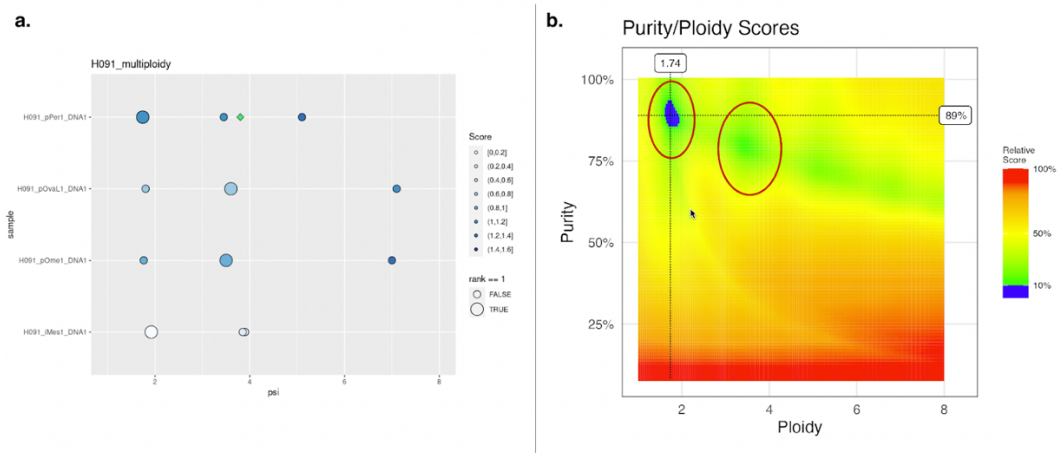


Fig 4: Ploidy correction. A) Multiploidy plot for patient H091: peritoneum and mesothelium samples result in a ploidy of 2 while ovary and omentum of about 4. The peritoneum sample has a high score and has been chosen to be corrected to a ploidy of about 4 and it is indicated by the green square. B) Sunrise plot for sample pPer1 of H091 patient: red circles highlight the chosen ploidy (blue dot) and the corrected ploidy (green minimum).

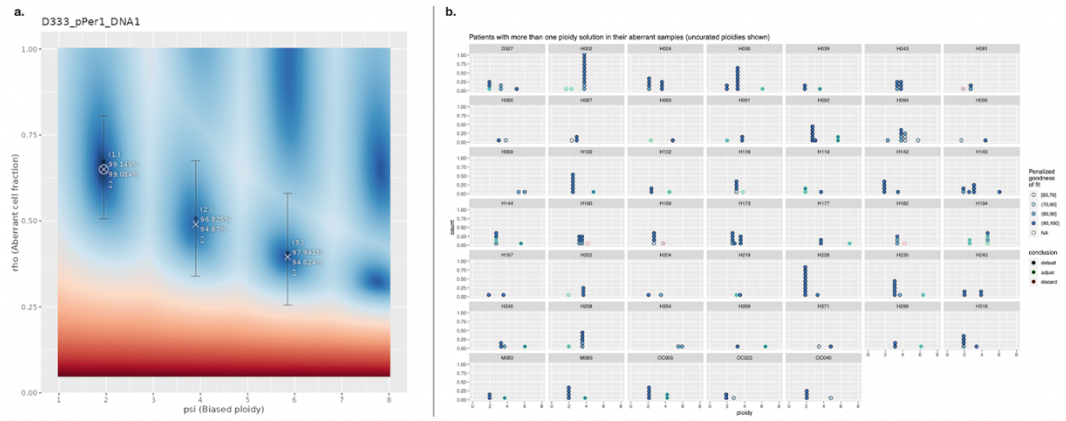


Fig 5: ASCAT sunrise plot and multiploidy. A) Sunrise plot for pPer1 sample of D333 patient: all grid combinations of ploidy/purity are plotted as function of ploidy and purity. The color scale represents the goodness of fit transformed as percentage (the higher the better). Red values represent low scores while dark blue values represent high scores. The best solutions are indicated with a cross together with the ranking, the goodness of fit and the penalized goodness of fit. Error bars are also reported for these three solutions. B) Multiploidy plot: Each patient for which samples have different ploidy estimates are extracted and plotted. Each section represents a patient and inside it all best solutions for the samples belonging to that patient are shown as circles. The color represents the goodness of fit while the outline of the dots indicates in black if the solution is the default one (calculated by ASCAT), in green if adjusted (manually curated) and in red if discarded.

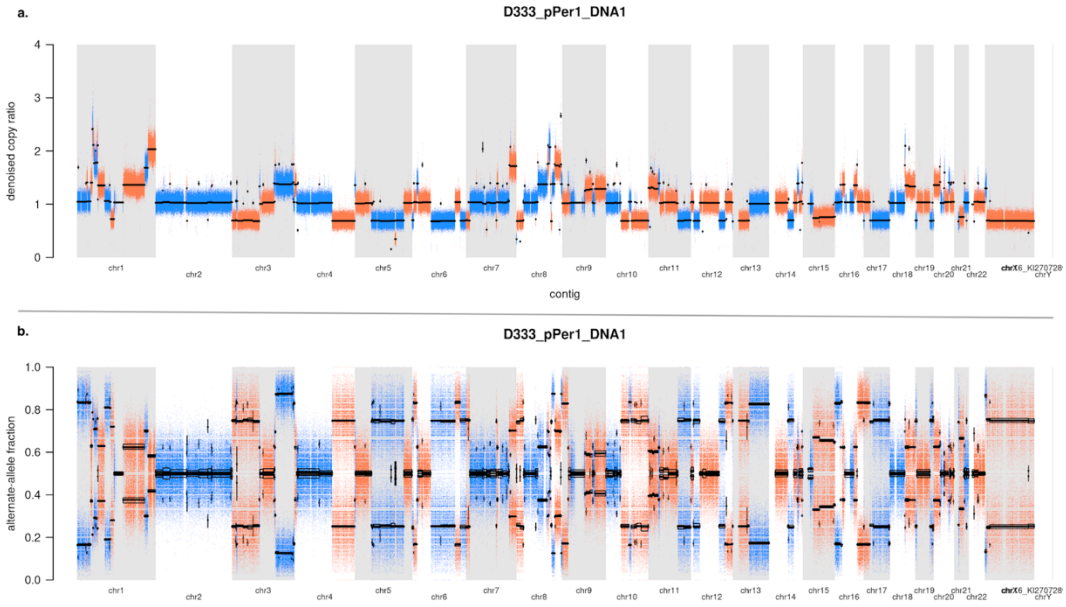


Fig 6: D333 pPer1 CN and BAF plots. A) Distribution of the denoised value of copy number for each segment along the genome. The colored shadows represent the single values and the color change represent the shift from a segment to its adjacent. The mean copy ratio for each segment is shown as a black line. B) BAF (alternate-allele fraction) values for each segment along the genome. The shadows and color shift have the same meaning as the LogR.

GATK-ASCAT results, not produced by this project, are produced in a similar way as Purple but organized in different files. The segmentation of all samples is supplied in a unique file called combinedAscatSegments.csv while all the possible candidates for the estimation of purity and ploidy are stored in combinedAscatEstimates.
As far as plots concern, ASCAT returns in a similar way to Purple, sunrise plots for each sample that allow checking the purity ploidy values (Fig5a). The plots work in a similar way, the only difference in the color scale and the score used for the estimation: in Purple the best score is the lowest, in

ASCAT the best goodness of fit is the highest. The top three estimates are reported together with the error bars and the value of the goodness of fit.
The multiploidy patients analysis have also been performed and the plots produced are built exactly in the same way (Fig5b). The only difference is that manual curated ploidies have been highlighted in green while the discarded ones in red. The color scale used for the ploidy points is the same but since the score goes in opposite directions, in the ASCAT case the darker means the best. The number of patients for which the sample presents multiple values of ploidy is different with respect to Purple but this is related to the differences in the overall pipelines and in the estimation of purity and ploidy.
In addition to these images, GATK also provides the representation of the distribution of copy ratio and BAF along the genome for each sample (Fig6). These plots permit the analysis and description of the segmentation.
The copy ratio plot (Fig6a) represents the number of copies for each segment with respect to the matched-normal sample. A value of 1 represents the normal situations, meaning a ploidy of 2, while a higher value represents an increase in the number of copies (amplification) and a lower value a decrease (deletion). The blue and orange shadows are composed by the single points (one of each read), for which the copy number has been calculated.
BAF is the frequency of the alternate allele, the allele with the minor frequency, with respect to the normal allele. The plot representing it (Fig6b) allows to evaluate the presence of a copy number variation in the presence of an heterogeneous sample. Since it is a fraction, BAF values can range from 0 to 1. Areas of homozygosity have BAF of 0 or 1; normal diploid regions have BAF of 0, 0.5, or 1. Homozygous deletions have no detectable signal so the calculated BAF appears as noise, but it is not present in the plot since it has been denoised. Copy number gains or losses cause the fraction to vary from these values, thus these areas of allelic imbalance show intermediate values.

The comparison of GATK-ASCAT and HMF Pipelines for the estimation of the purity and the ploidy

Given the difference of the algorithms used for the estimation of the purity and the ploidy (ASCAT and Purple), the first check is the comparison of the ploidies.
Samples are represented as points in a scatter plot (Fig7) where the axes stand for Purple and ASCAT ploidies (x and y axis respectively). The blue line shows the region where both ploidies are equal, and it is clear from the plot that most of the points lie in it. This means that the two algorithms are in agreement and consistent as far as ploidies concern.
Nevertheless, samples with a difference in ploidies is higher than 1.5 are highlighted in red and looking to the sunrise plots from both the pipelines they fall in three categories:

- Tools estimate differently the ploidies (ex. H201, H194 and D237)
- Purple estimate seems to be unreliable (H119)
- Low purity (ex. H089, OC040, OC038)

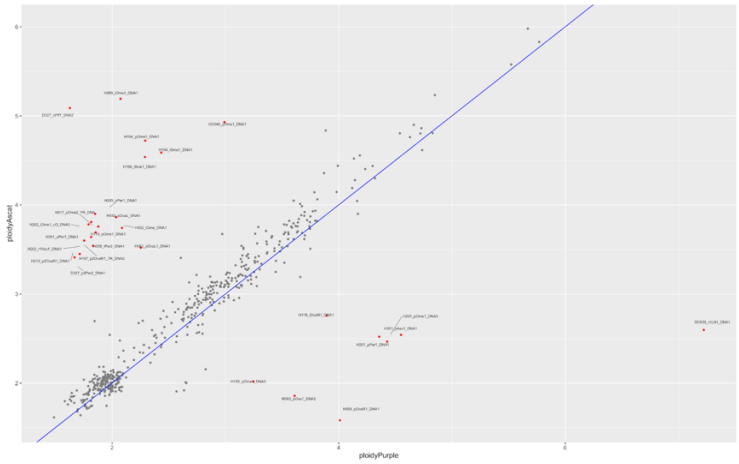


Fig 7: ploidy comparison between pipelines: best estimates of ploidy for each sample from GATK-ASCAT and HMF pipeline are represented in a scatter plot (x axis Purple ploidy, y axis ASCAT ploidy). Samples differing in the ploidy estimation for a value higher that 1.5 are highlighted in red with their ID. The blue line represents the bisector, thus the region where points have the same value for x and y axes.

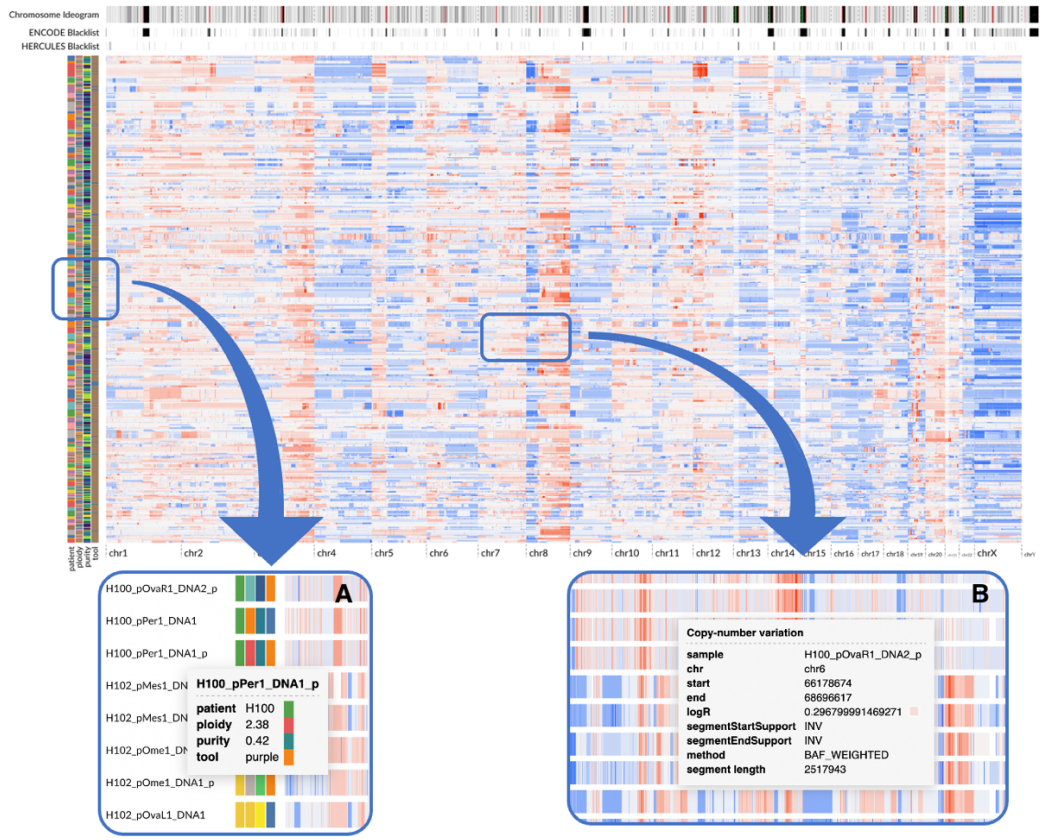


Fig 8: GenomeSpy depiction of the segmentation of all samples from HMF and GATK-ASCAT pipelines. In the main picture starting from the top the chromosomal cytobands are represented, both ENCODE and HERCULES blacklists and the segmentation. In the left part for each sample, patient, purity, ploidy and used tool metadata are displayed. The different colors for metadata bars indicate different values. In the segmentation part, the different color of the segments specifies the value of the logR (red means amplification, blue means deletion). Zoom A) Going with the pointer over a sample, all values for metadata are shown in a window. IDs are different for the pipelines: HMF pipeline samples have a "_p" at the end of the ID. Zoom B) Going with the pointer over a segment, different values from the segmentation are exhibited: sample name, chromosome, start and end position, logR value, start and end support of the segment, detection method of the segment and segment length.

The next analysis is done on the segmentation provided by the two pipelines. The best method for its visualization is a plot built in GenomeSpy in json format (Fig8). Starting from the top of the top of the plot there is a bar representing the chromosome ideogram with the cytobands given by the Giemsa staining. Below it, there are the two blacklists, which are comprehensive sets of regions that have anomalous, unstructured or high signal in high throughput sequencing experiments independent of cell line or experiment. Removal of blacklists is an essential quality measure for genomic analysis. ENCODE blacklist is the most used and built by the ENCODE project consortium (used by Purple), while the DECIDER blacklist has been built by GATK from the samples used. The main plot shows the copy number profile for each sample from both GATK-ASCAT pipeline (normal sample name) and HMF pipeline (sample name added with "_p"). The color of the segments represents the value of the logR and positioning the mouse on one segment it is also possible to have more information as shown in the Fig6B zoom, like the position of the segment, the precise value of the logR, support of star and end point (centromere, telomere, inversion, deletion…) and the method used for establishing the segment (BAF-weighted, SV…). In the left part different metadata are reported for each sample: patient, ploidy estimation, purity estimation and the pipeline used.

Fig 9: Segmentation details. A) Difference in the number of segments in the same samples between GATK-ASCAT pipeline and HMF pipeline. HMF samples ("_p" IDs) show a clear increase in the number of segments. B) Low purity samples segment recovery: in samples with a low value of purity (such as H033_pLN1) HMF pipeline is able to recover some segments (light blue and light red) that are missed by GATK-ASCAT pipeline (grey bar). C) Short segments detection: HMF pipeline recovers short segments (highlighted red small bar) supported by structural variants (SGL start and end support).

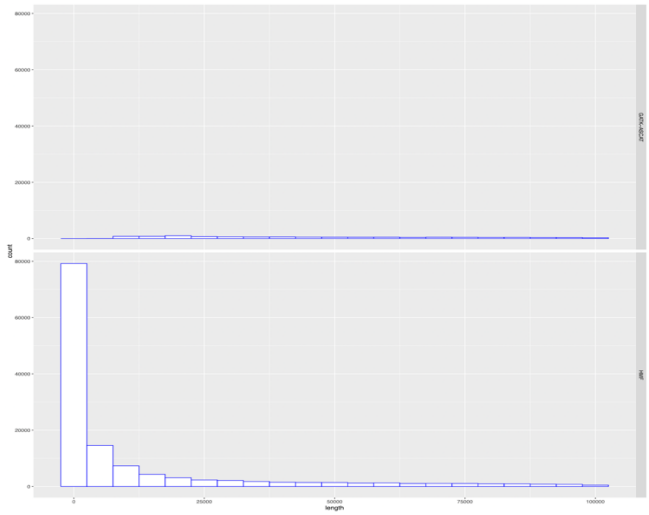


Fig 10: Histogram showing segment length distribution for GATK-ASCAT and HMF pipelines. Segments shorted than 100kb have been extracted for both pipelines and plotted in a histogram. In the upper plot GATK-ASCAT distribution is shown, while in the lower plot HMF distribution. Bin width for both histograms is 5kb and for each bin the height represents the count of the segment with a length in the range of the bin.

Looking in detail at the single samples it is possible to see that copy number profiles obtained with Purple are much more segmented, especially in very small segments as it is shown in Fig9a. These short segments mostly recovered thanks to the use of SVs breakpoints and are missed by GATK because the window size for detecting changes in the copy ratio is higher than the total length of the segments. Fig9c shows one of these cases: the red segment is only 20bp long and it is reported to be an inversion (method) with both ends supported by a single-breakend SV (SGL). Exploring in detail the plot, it is evident that it is not an isolated case, but a lot of segments are recovered from Purple and this fact is also confirmed by the histogram in Fig10. The distribution of the number of segments shorter than 100kb is shown both for Purple (bottom plot) and ASCAT (top plot). The total number of segments from 0 to 5kb reaches almost 80 thousand in Purple while in ASCAT is very low (less than 2500).

GenomeSpy plot also depicts the recovery of some segments in samples with a purity close to 0 in Purple data, which ASCAT is not able to do (Fig9b). In sample H033_pNL1, whose purity is estimated to be 0.08, the ASCAT bar has only very long segments with a logR very close to 0, while Purple bar shows some segments that are also present in H033_pOvaR1 sample which has an higher purity. It means that Purple is more sensible in detecting segments also in low purity data, even if some segments are not reliable.

A further confirmation of the correctness of segmentation can be obtained with another visualization in GenomeSpy in which for a single sample the segments from both the pipelines are reported together with the logR of the single points (Fig11). The total amount of points has been reduced by 50% to not create too much crowding. Orange segments and points represent HMF pipeline output while green points and segments represent GATK-ASCAT pipeline output. In the zoom B it is shown that for each point it is

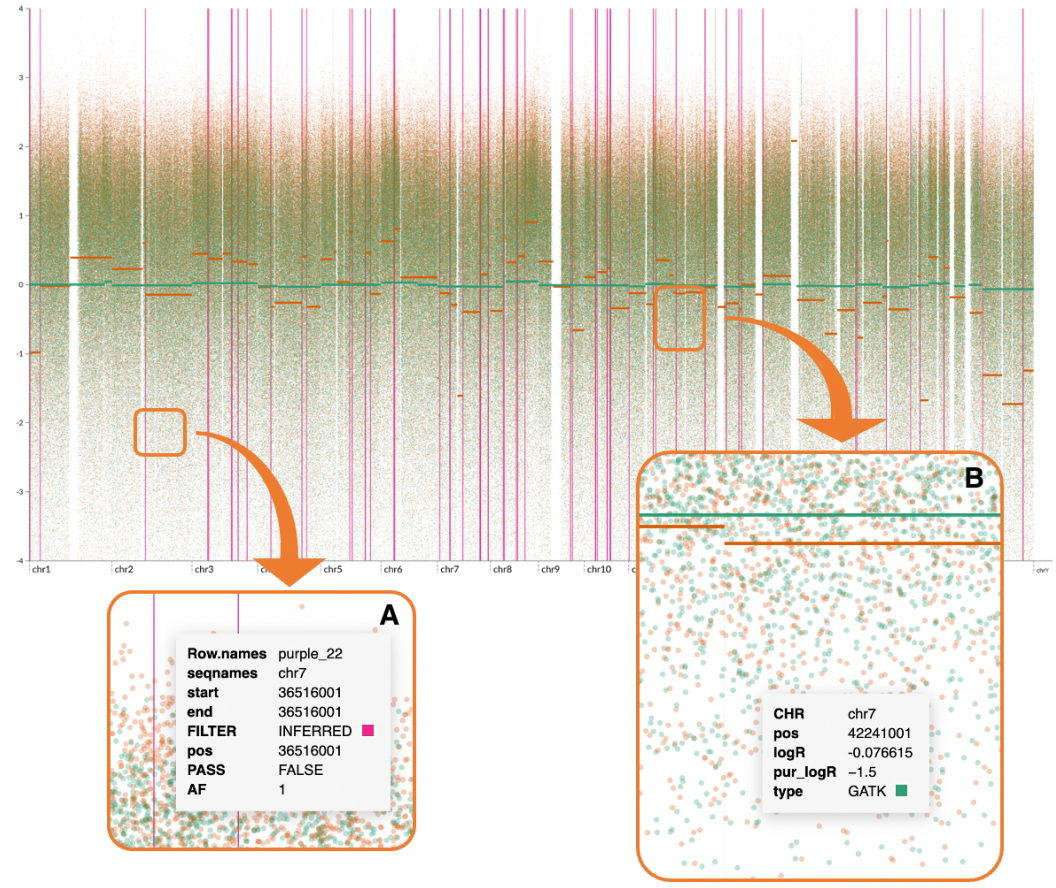


Fig 11: GenomeSpy representation of the segmentations of a sample from GATK-ASCAT and HMF pipelines. Green points and segments denote GATK-ASCAT pipeline while orange points and segments HMF pipeline. y axes indicate the logR while x axes the position in the genome. For each position in the genome the logR has been calculated and points symbolize it. Zoom B shows that pointing one of them, position, logR, purified logR and pipeline type are shown. Segments correspond to the result of the segmentation process. Vertical lines stand for the structural variants recognized by GRIDSS. Zoom A shows the details retrieved by pointing to one structural variant: SV ID, position, filter (INFERRED from BAF and read depth or PASS from a SV) and allele frequency (AF).

possible to retrieve the purified logR, the position in the genome and the pipeline type. The vertical lines represent the structural variants, or the breakpoints used by Purple to establish the segments. Gray bars symbolize breakpoints inferred from BAF differences while pink lines from structural variants. Each line gives the position in the genome of the breakpoint, the reference in the GRIDSS output file, the type of breakpoint and the allele frequency.

Looking in detail at sample H110_pOvaR1 segmentation (Fig12), in a zoom of the overall plot, it is clear how GATK-ASCAT misses some segments which are instead identified by HMF pipeline.

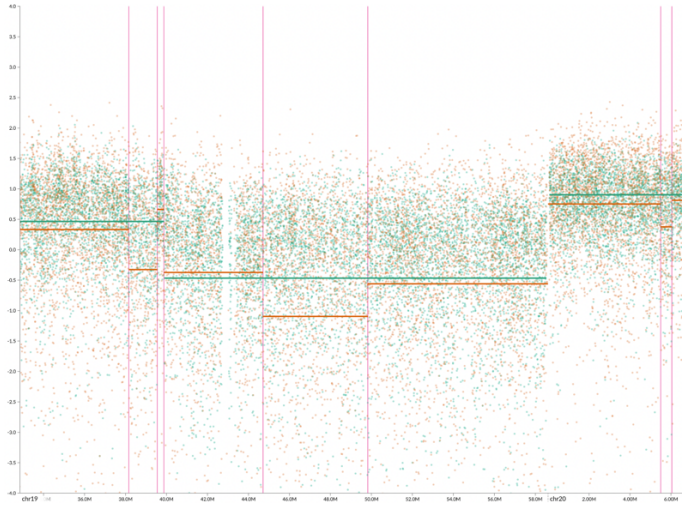


Fig 12: Detail showing discordance between GATK-ASCAT and HMF pipeline in H110 pOvaR1 sample. GATK-SCAT pipeline (green) misses some segments detected by HMF pipeline (orange). The evidence for the presence of the segments is provided by structural variants (vertical lines).

RESULTS FROM SIGNATURE ANALYSIS

Quantification of COSMIC signatures

Cohort sample copy number profiles have been processed using SigProfilerMatrixGenerator, which provides their summarization in a S x 48 matrix, where S is the number of samples and 48 are the classes derived from the combination of the heterozygosity state, TCN and segment size features. The final matrix is called $\boldsymbol{M}$.
COSMIC signatures have been downloaded and quantified in HGSC samples using the three mentioned methods: SigProfiler, non-Poisson model and Poisson model. The results consist of matrices ($\boldsymbol{H}$) in which the activity related to each signature is calculated for each sample.

SigProfiler decomposition module has been used with its cosmic_fit function which fits the provided signatures by using the NNLS (Non-Negative Least Squares) algorithm, whose generalization is NMF. The goal of the algorithm is solving a linear least squares problem with the constraint of non-negativity on the solution. It means that the equation Xβ = y must be fit ensuring that 0. Given the model function $y = f(x, \beta)$, where x are independent data points and β the parameters, β values are found so that they are ≥0 and the equation

$$f(x, \beta) = \sum_{j=1}^{n} \beta_j \varphi_j(x)$$

is respected. To do this, it is necessary to find the minimal possible value of the sum of squares of the residuals:

$$r_i(\beta) = y_i - f(x_i, \beta)$$
$$S(\beta) = \sum r_i^2(\beta)$$

To select for each sample how many signatures to attribute, a forward-backward process is performed. For each sample, the maximal signature exposure is taken as a starting point and then a second signature exposure is added.

To decide whether to keep it or not, the cosine similarity is calculated between the two vectors:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Where A is the vector with only one signature and B the vector with two signatures. Cosine similarity is a metric used to measure the similarity between two or more vectors and consists in the cosine of the angle between two or more vectors (typically non-zero and within an inner product space). It is described as the division between the dot product of vectors and the product of the euclidean norms (magnitude) of each vector. It is bound by a constrained range of 0 and 1.
The second signature is added if the cosine similarity is higher than an arbitrary threshold (in this case 0.1).
The backward part instead does the opposite process, subtracting one signature exposure per time and checking if the cosine similarity decreases. If it decreases over a certain threshold (in this case 0.01), the signature is removed.
In the end what is done is assigning the attribution that gives the best cosine similarity between the input sample vector and the reconstructed sample vector [26].

The other two methods use the same strategy but apply different functions for the calculation of the exposures and the selection of the number of signatures. Both have a forward-backward approach like SigProfiler.
The Poisson model uses the Kulleback-Leibler divergence for the estimation of the activities of the signatures in the samples. It measures the information lost considering the probability distribution p and the approximating distribution q. The sum of the difference of their log values represents the divergence:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \cdot (\log p((x_i) - logq(x_i))$$

This measure of difference between two distributions can be used, as in this case, for an optimization problem. It is possible to choose the values of $p(x_i)$ so that the KL divergence is minimized to have as little information loss as possible.
This measure is used in the context of NMF for calculating the distance between the target matrix and its NMF estimates and minimizing it.
The number of signatures for each sample in the Poisson model is chosen using BIC as the selection function. It is an index used to choose between two or more alternative models. It is defined as:

$$BIC = klog(n) - 2\log(L(\theta))$$

where n is the number of data points, k the number of parameters and $L(\theta)$ the likelihood. The model with the lowest BIC is considered the best. The likelihood is calculated as the density of the Poisson distribution.

In the Non-Poisson model, the method for calculating the exposures of the activities is the same used by SigProfiler, NNLS, while the function to choose the number of signatures is the cosine similarity.

The best method has been evaluated by calculating the cosine similarity between the activities of the signatures $\boldsymbol{H}$ and the product of the reference

signature matrix ***W*** and ***M***. This metrics allows to inspect how much the calculated activities explain the data because its multiplication with the reference signatures produces a matrix similar to ***M***. This matrix is compared to ***M*** and if the cosine similarity between each corresponding column is similar, it means that the signatures are able to properly and fully explain the used segmentations.

The average cosine similarity for the three methods is reported in Table2. There is not a huge difference between the three methods, but it is still possible to notice that the non-Poisson model offers the best result.

Table 2: cosine similarity for COSIM CN signatures fitting (sig = SigProfiler, p=Poisson, no=non-Poisson)

type	cos_sim
act_sig	0.7818696821658829
act_p	0.7850401502222072
act_np	0.8102404485360392

A better inspection of the data is provided by a GenomeSpy visualization (Fig13) which reports as metadata all the mutational signatures (SBS, DBS, ID) and the copy number signatures. Some signatures are represented as a yellow bar because their quantification is 0 in all samples.

By ordering the samples by SBS3 signature, which is associated with HRD [14], and inspecting the distribution of the values of the other signatures, it is clearly visible that the highly segmented copy number profiles are associated with high levels of SBS3, while there seems not to be a correlation with the copy number signatures. In particular, CN17 signature, which has been associated with HRD [19], seems to show an opposite trend with respect to SBS3.

The confirmation of these observations comes from the cosine similarity calculated between the vector of SBS3 activities and the vector of the others copy number signatures.

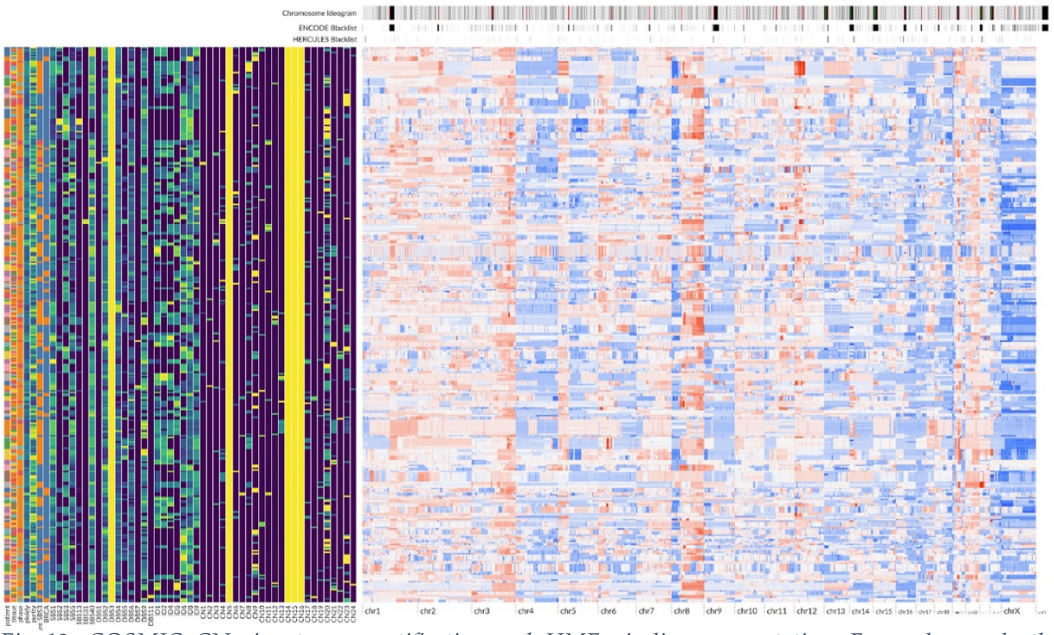


Fig 13: COSMIC CN signature quantification and HMF pipeline segmentation. For each sample the segmentation is reported in the left part and in the right part the activities of the mutational signatures, copy number signatures, ploidy, purity, tissue site and treatment phase. Yellow bars indicate a value of zero for all samples while in the other bars the color scale goes from 0 (purple) to 1 (light green).

Cosine similarity is used as a measure of association between the signatures instead of other correlation metrics (Pearson, Spearman or Kendall correlation). Correlation metrics are not used because of two main reasons:

1. Reciprocal dependence: each patient has more samples; thus these samples will be correlated because they share the genome. It can be demonstrated by calculating the correlation between each couple of samples in the ***H*** matrix and the p-value of each correlation. The result is that each couple shows a positive correlation with a significant p-value and as a consequence, each measure of correlation between two signature vectors will not work.
2. Skewness of marginals: activity values range from 0 to 1 but their distribution is not normal and there is a high number of samples with the activity equal to 0, as seen in the histogram of the signature SBS3 (Fig14a). For this reason, it is not possible to find any kind of correlation between two signatures, even if there is clearly a relationship. One example is provided by the scatter plot in Fig14b for the signatures SBS3 and ID1 which are negatively correlated [14].

Cosine similarity calculated between SBS3 and each copy number signatures clearly shows no association with any signature (Table3). NA values are generated because all samples have 0 activity for that signature, thus the operation cannot be completed.

Table 3: cosine similarity between COSMIC CN signatures and SBS3 mutational signature.

CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	CN9	CN10	CN11	CN12
0,0000	0,0309	0,0294	0,1671	NA	0,1778	0,1074	0,0448	0,4795	0,1357	0,1109	0,0333
CN13	**CN14**	**CN15**	**CN16**	**CN17**	**CN18**	**CN19**	**CN20**	**CN21**	**CN22**	**CN23**	**CN24**
0,0000	NA	NA	NA	0,0834	0,0000	0,0917	0,5305	0,0218	0,2320	0,4484	0,0000

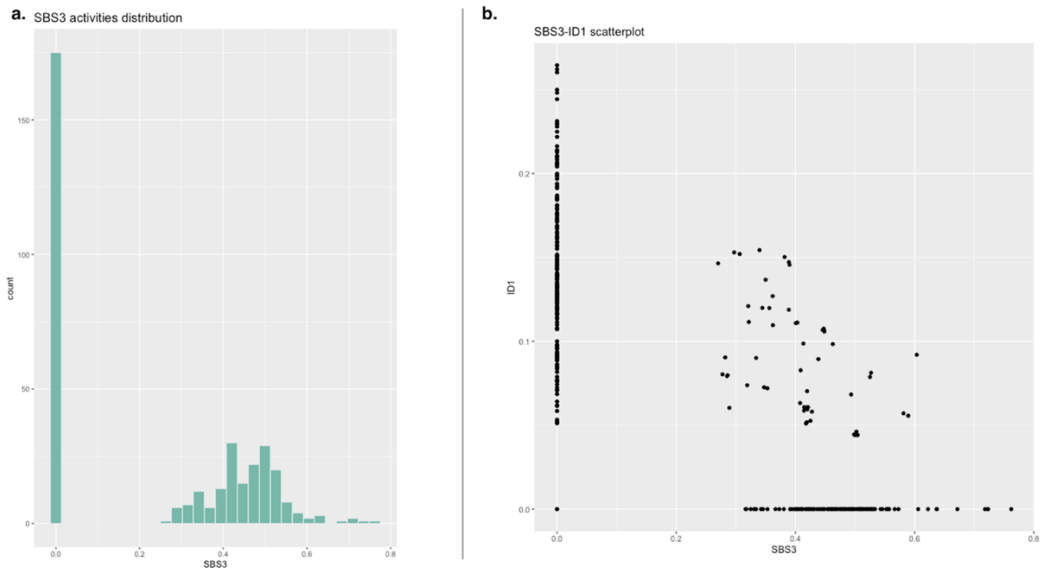


Fig 14: Analysis of signatures activities. A) SBS3 histogram displaying the number of samples for each range of activity estimation (bin width 0.02). B) Scatter plot of all samples for signatures ID1 and SBS3. Black dots represent samples as a function of the activity of the two signatures (SBS3 x axis, ID1 y axis).

Exposure calculation in long-segments dataset

COSMIC signatures have been extracted from data processed with ASCAT, which produces copy number profiles very different from the profiles generated with Purple as already seen. Purple profiles have been seen to be more sensitive to very short segments (Fig8), thus it is possible that COSMIC signatures do not fully explain copy number profiles because of the abundance of these small segments.

To prove this point, Purple copy number profiles have been modified by removing all small segments (smaller than 10kb) and merging the remaining ones. The profiles so obtained underwent the same procedure for summarization and quantification of COSMIC signatures. The amount of information explained by these new exposures has been evaluated in the same manner through cosine similarity.

Table 4: cosine similarity for COSMIC CN signatures in the original dataset and dataset with long segments.

type	cos_sim
act_sig	0.7818696821658829
act_p	0.7850401502222072
act_np	0.8102404485360392
act_sig_long	0.804820651254986
act_p_long	0.8072188457573296
act_np_long	0.8310268871036831

As reported in Table4, there are two points of difference between each method applied to the original dataset and the dataset with long segments., meaning that the situation improved but not significantly. Moreover, the problems in the association of copy number with mutational signatures still remain.

De novo extraction of copy number signatures

COSMIC copy number signatures did not give promising outcomes, as a consequence, de novo extraction has been performed directly on Purple copy number profiles to obtain signatures that explain as much as possible those profiles.

The method used follows exactly the state-of-the-art procedure as in [19], except for the first part, where the code of SigProfilerMatrixGenerator has been modified. The feature concerning the segment length has been added with one class: the previous first class 0-100kb has been split in 0-25kb and 25kb-100kb classes. The threshold of 25kb has been arbitrarily chosen looking at the histogram in Fig10 so that there is one new class catching all small segments missed by ASCAT.

Not all samples have been used for the de novo extraction because, since there are different numbers of samples per patient, patients with a higher number of them would have a stronger impact on the extraction to the detriment of the others. Nevertheless, the biological variation given by the treatment phase must be kept, thus the selection performed included the highest purity sample of each treatment phase in each patient.

Copy number profiles of the selected samples have been summarized using the new 58 features and then used for the de novo extraction performed by SigProfilerExtractor. All parameters used for COSMIC signatures extraction have been maintained.

The resulting 15 signatures have been first decomposed by SigProfilerAssignment and then used for the quantification of the exposures in the complete dataset of samples. The three methods previously used have been repeated also in this case and the amount of information explained by each method is calculated again with cosine similarity. Table 5 shows that there is an improvement of four points using SigProfiler, three with Poisson model and one using non-Poisson model. Even if the increase is not so stunning, a more in-depth analysis has been conducted.

Table 5: cosine similarity for de novo extracted copy number signatures.

type	cos_sim
act_sig	0.8283194438450688
act_p	0.8188304688276209
act_np	0.8239049661964596

Genome Spy visualization includes all the new signatures (SBS58A-SBA58O) and some clinical and mutational data like the presence or absence of CDK12, MYC family or CCNL mutations, SBS3 signature, PFI, and others (Fig13a).
Ordering the samples according to SBS3 allows to see a similar variation also in SBS58B and SBS58D, giving the hint they could be associated with HRD (Fig13b).
The confirmation of this association comes from cosine similarity, which has been calculated with all signatures and reported in Table 6.

Table 6: Cosine similarity between de novo extracted copy number signatures and SBS3.

SBS58A	SBS58B	SBS58C	SBS58D	SBS58E	SBS58F	SBS58G	SBS58H
0,5525	0,7722	0,5393	0,5695	0,0258	0,5230	0,6657	0,1786
SBS58I	**SBS58J**	**SBS58K**	**SBS58L**	**SBS58M**	**SBS58N**	**SBS58O**	
NA	0,2384	0,3673	0,2687	0,0859	0,0647	0,0053	

In particular, the association with signature SBS58B and SBS58G is quite important and could explain HRD phenotype exactly like SBS3.

De novo signatures have been visualized through activity plots (Fig16) in which it is possible to see which features each one mainly explains. In particular, taking the highest peaks of each one and comparing them with the activities of COSMIC signature it is evident that some of them can be associated with them and in particular can be explained by the same process (Table7). Nevertheless, these associations must be proved with other methods.

Signature	Features	Possible cause
SBS58A	Hom del short and LOH 1 short	short deletions both hom and LOH
SBS58B	short segments LOH 1,2,3-4 HET 2, 3-4	short deletions followed/following 1xWGD
SBS58C	Long LOH 1 and longer HET 2	chromosomal instability on a diploid background
SBS58D	short LOH 1,2 all HET 3-4, 5-8	tandem duplication
SBS58E	medium-high LOH 1,2 and medium HET 2, 3-4	chr LOH
SBS58	low fragments HET 3-4, 5-8 and LOH 3-4, 5-8	focal LOH and 1xWGD
SBS58G	long LOH 2 and HET 3-4	chr LOG and 1xWGD
SBS58H	short and long LOH 2, 3-4(a lot), 5-8	
SBS58I	long Hom del and long LOH 1	long deletions
SBS58J	very long LOH 1, a bit long HET 2	chr LOH?
SBS58K	medium LOH 3-4, 5-8, all HET 5-8	chr LOH and 2xWGD
SBS58L	long LOH2 , HET 3-4	chr LOH and 1xWGD
SBS58M	short HET 9+	chromotripsis associated amplification
SBS58N	short LOH 5-8, 9+	
SBS58O	LOH 1 long	LOH of long fragments

Table 7: Description of de novo extracted copy number signatures with main features represented and possible explanation of the cause.

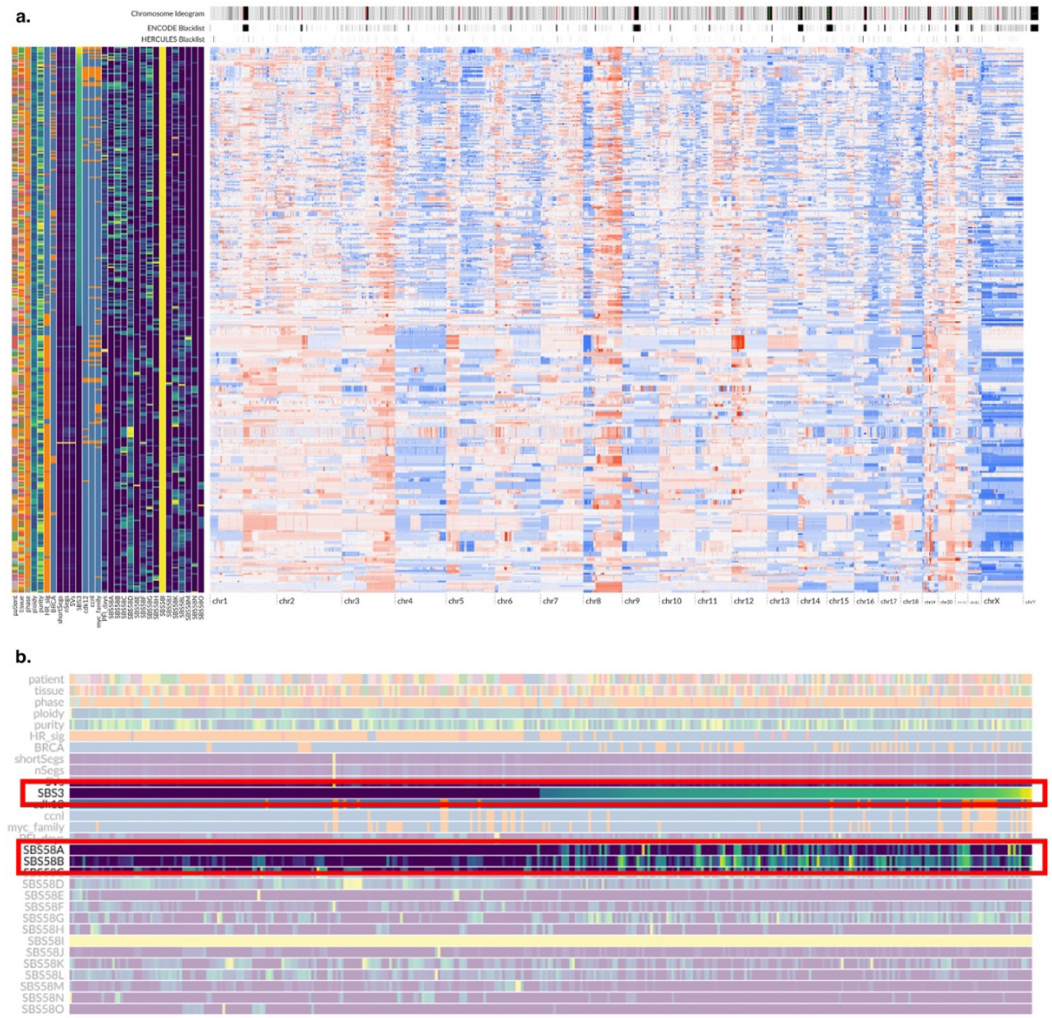


Fig 15: A) de novo extracted signatures values compared to the segmentation. In the left part metadata are described: patient, purity, ploidy, tissue, patient, HR status (HRD/HRP), CDK12 mutation, Myc mutation, SBS3 signature, SBS58 signatures. In the right part the segmentation for each sample is shown. Samples have been ordered according to the value of SBS3 activity. B) Zoom of the metadata bar highlighting the similar behavior of SBS3 and SBS58A and SBS58B.

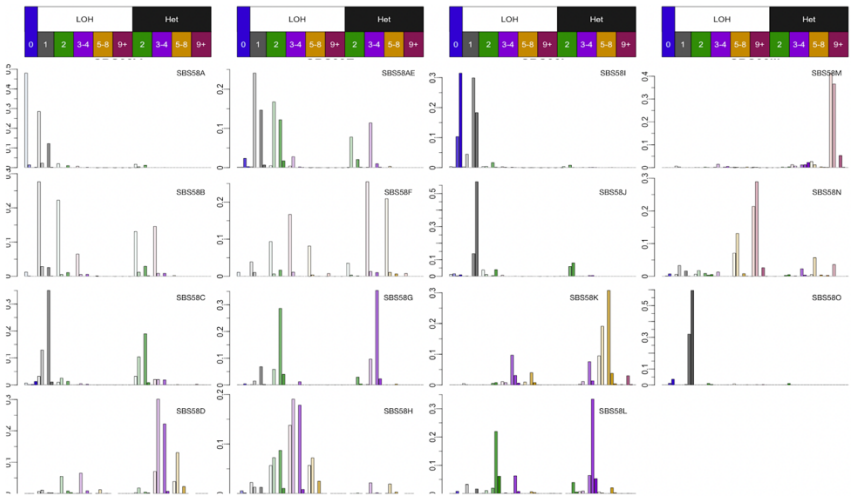


Fig 16: activity plots of each de novo extracted signature. Each bar plot indicates the percentage of contribution of each feature to the signature. In the x axis all the CN features are reported (shown in the colored boxes over each column of bar plots) including copy number status (LOH, Heterozigous or Homozygous deletion), total copy number value and segment length. The color of the bars reflects the belonging category of total copy number value (0, 1, 2, 3-4, 5-8, 9+).

CHAPTER 4: DISCUSSION

The presented work shows the implementation of a new pipeline for the detection of somatic copy number variations in a cohort of patients diagnosed with high-grade serous ovarian cancer. Data extracted from the samples and analyzed with it have been then used for the research of recurrent patterns of copy number variations by quantifying COSMIC signatures first and then extracting a set of 15 de novo signatures which better explain used data.

The pipeline produces results that show an improvement in the segmentation profiles with respect to GATK-ASCAT as it is demonstrated by GenomeSpy representations. GATK fails to catch segments shorter than 1kb because for detecting variations in read depth, it uses a window size of exactly 1kb. Purple instead, exploits GRIDSS detection of the structural variants, which are prioritized in the establishment of segments edges. Since GRIDSS considers all samples coming from the same patient as a unique sample, the power of detection is increased and also variants which are present in small amounts are detected. This improvement can be observed especially in the low purity samples, where some segments can be detected despite the low amount of cancer cells (Fig9b).
Moreover, it makes use of different read features and local assembly for variant detection, which permit the algorithm to identify variants that are few bases far away.
Besides these advantages, one point to be verified remains the ploidy estimation, for which the two pipelines differ by more than 1.5 for the ploidy value in some samples. The validation of the ploidy of these samples should be performed in vitro by a Fluorescent In Situ Hybridization (FISH) in specific portions of the genome where there is a detected concordance in the segmentation. The technique is a laboratory technique used to detect and locate a specific DNA sequence on a chromosome. In this technique, the full set of chromosomes from an individual is affixed to a glass slide and then exposed to a probe, which is a small piece of purified DNA tagged with a fluorescent dye. The fluorescently labeled probe finds and then binds to its matching sequence within the set of chromosomes. With the use of a special microscope, the chromosome and sub-chromosomal location where the fluorescent probe is bound can be seen and quantified in its number of copies [28].

The visualization in Fig15a, in which samples have been ordered second to the HRD signature SBS3, displays a clear difference in the segmentation profiles of HRD samples (in the top part) which are visibly characterized by an highly fragmented genome, and HRP samples (in the bottom part) whose genomes are more uniform. This feature was not possible to be identified from GATK-ASCAT segmentation, representing a clear improvement.
Other refinements could be obtained by building up a panel of normals from samples of the used dataset and not a general panel, improving the estimation of the purity using TP53 mutation and adding the position of oncogenes, tumor suppressors and point mutations to the GenomeSpy visualization.

As previously explained, the PoN has the function to normalize all the artifacts derived from the sequencing technique or other manipulations to improve the variant calling. It must be made of normal tissue to avoid normalizing the real variants. As a result, the most important selection criteria for choosing normals to include in any PON are the technical properties of how the data was generated. It's very important to use normals that are as technically similar as possible to the tumor (same exome or genome preparation methods, sequencing technology and so on). The PoN used comes from a Dutch population and has been created by Hartwig Medical Foundation but providing a tailored PoN could ameliorate the analysis.
The BAF of the truncal *TP53* mutation can be used for adjusting the purity calculated by tools like ASCAT and GRIDSS transforming it into a *TP53*-based purity, thus as additional evidence for the optimal ploidy/purity selection. By using the total copy number and *TP53* BAF, purity can be approximated as:

$$purity = \frac{2}{(CN/VAF) - (CN - 2)}$$

The gene *TP53* encodes the tumor suppressor protein p53 which regulates cell division by keeping cells from proliferating too fast or uncontrollably through senescence induction, cell cycle arrest, DNA repair or changes in metabolism. This mutation is used because HGSC has the highest frequency of p53 mutation of any solid cancer, approximately 97% and it has been recognized as a truncal mutation, which is a mutation that is present at the trunk of the cancer evolutionary tree [29]. The frequency of TP53 mutation, together with the inferred copy number of the tumor, can be associated directly to the number of cancer cells present in the sample and hence to the purity.
The *TP53*-based purity can be compared to the calculated purity and if it is too different the model for estimation of the purity and ploidy is rejected and a better model needs to be selected.
The last improvement can be done in the visualization by adding the position of relevant genes in order to see segmentation differences directly in those genes, and by adding the SNVs extracted with additional analysis and used also in the HMF pipeline so that it is possible to notice if there are recurrences in certain patterns of segmentation and point mutations.

Limitations for this pipeline include the poor use of HMF tools and the reliability of low purity samples. Since Purple is still not widely used in the state-of-the-art pipelines by the scientific community, a lot of tools do not consider it. This implies that the output must be modified to a format compatible with these tools like ASCAT, SEQUENZA or ABSOLUTE. Something missing for example is the estimation of the LogR and LOH, which must be calculated posteriorly.
As far as low purity samples are concerned, the HMF pipeline is able to extract some segments, which is something that the GATK-ASCAT pipeline is not able to do. Despite this, it is clear that some segments are totally improbable because not matching with samples from the same patient (copy number value completely different) or because present in regions excluded by DECIDER blacklist (Fig9b). This observation raises the

question of how much to trust these segments and if it is necessary to provide an additional step of filtering for the most improbable segments.

The progresses that can arise from this new pipeline can be seen especially in the field of tumor evolution and in the identification of new targettable variations such as structural variants affecting tumorigenic genes.
One of the main questions regarding HGSC is if patients can be divided into groups according to their genomic features and if this stratification allows to discover new targets for treatment and diagnosis. One step forward the accomplishment of this task consists in the study of the tumor evolution, which has been proved to be effective using SNP data, but results challenging with sCNAs. The identification of subclones and the understanding of the dynamics causing one to evolve in another require the signal to be deconvoluted. The deconvolution of the signal coming from samples is complicated as both the sCNAs and the proportion of cells originating from each clone in the mixture are unknown [30]. The use of the joint calling for samples coming from the same patients and thus the identification of variants present only in some subclones reasonably can help the analysis. Moreover, the extraction and prioritization of structural variants for the segmentation instead of considering the differences in read depth alone could further enable to unravel the mutational processes at the base of the observed copy number patterns.
Moreover, the deeper resolution in the segmentation can also determine and identify with higher precision the position of structural variants. Structural variants have a fundamental role as driver events for initiating tumorigenesis and to shape the tumor genomes, that's why their improved identification in cancer can lead to more targeted and effective treatment options as well as advance our basic understanding of the disease and its progression [31].

The stratification of the patients according to sCNAs has been approached by using the copy-number signatures published in COSMIC [19]. The activity of the 24 identified signatures have been quantified in each sample using different statistical methods, the reference tool and two other methods previously used in the extraction of the exposures of COSMIC SNV signatures. The implemented methods have been introduced because computationally faster and because since the non-Poisson model has been used for the extraction, it could have been more efficient also in the attribution.
None of the three procedures produced remarkable results in terms of cosine similarity, meaning that they are not able to properly explain the data. This is possible because the extraction of the signatures has been done in a pan-cancer context, using 33 cancer types that are very different in the genomic and cellular characteristics from HGSC [19]. Ovarian cancer samples represent 5% of the total but it has not been specified which subtype of ovarian cancer samples have been used. It is important to include all subtypes and a proper number of samples for each one because, as explained in the introduction, ovarian cancer subtypes are very different in terms of genetic characteristics, cellular features, progression and aggression [1]. Nevertheless, since HGSC is the most represented subtype of ovarian cancer, it is improbable the number of samples is not sufficient.

By the way, overall, it is possible that ovarian cancer has not been represented enough in the cohort used in the COSMIC study, causing a minor possibility of correct explanation of the used data.
In addition to that, it is also important to consider that the genome version used for COSMIC signatures extraction is different from the version used for preprocessing of the used data, meaning that positions of some features could be different.
Moreover, the type of data used in the COSMIC study come from different platforms consisting of WGS, whole-exome sequencing and SNP6-profiling-deriver copy number profiles but at the same time they claim that rearrangement signatures can only be derived exclusively from WGS data and cannot capture important prognostic information such as WGD [19]. That is possibly another explanation for the poor representation of HGSC WGS data by the signatures. Since WGS data have a higher resolution than whole-exome sequencing and SNP6 copy-number profiles, probably more detailed characteristics have not been caught by the COSMIC method.

One important feature that is looked for in HGSC genomes is the HRD/HRP phenotype. Indeed, HGSC is characterized by a high level of chromosomal instability which can be caused in most cases by homologous recombination repair (HRR) pathway deficiency. Germline *BRCA1/2* mutations and BRCA gene promoter methylation are well known causes of HRD, but other genetic abnormalities of the pathway can also cause this phenotype. Ovarian cancer with these alterations behaves in a similar way and this behavior is termed the "BRCAness" phenotype. Unrepaired DNA damage can result in accumulated mutations and unregulated cell division, and HRD is thus related to cancer susceptibility. Large amounts of DNA damage can lead to cell apoptosis but when only HRR is deficient, the activities of other DNA mechanisms can prohibit the accumulation of excessive DNA damage and apoptosis [6].
Indeed, the identification of BRCAness is particularly important because poly (ADP-ribose) polymerase (PARP) inhibitors in patients with HRD compromise another pathway of the Dna repair, the Base Excision Repair (BER).
Since two pathways involved in the DNA repair are not working, the cells cannot survive anymore and the treatment results in lethality for cancer cells. In other words, mutations occurring in one of two genes separately do not result in apoptosis, but the impairment of both genes simultaneously leads to cell death (synthetic lethality) [32].
Moreover, less research has been done on HRP patients, who generally have poorer outcomes and not that many treatment options. To help the development of efficient treatment for them, it is important to identify genetic variation in them to enhance drug development.
The clear identification of the HRD phenotype is thus important because it allows to provide an additional therapeutic strategy, but currently there is not a method that is clearly able to divide HRD from HRP patients. Some databases furnishing HRD scars have been proposed by different companies (for example, the Murrai test), but they have been shown to not identify proficiently HRD patients.
SBS3 signature has been clearly shown to be associated with HRD phenotype [14] because it is strongly associated with germline and somatic

BRCA1/2 mutations and *BRCA1* promoter methylation. Similarly, CN17 has been proved to be associated with mono- or bi-allelic losses of, *BRCA1/2*, *PALB2*, *FBXW7* and *CDK12*, and the strongest positive association with scarHRD scores (a score of genomic scars of homologous recombination deficiency) [19]. Nevertheless, when comparing the quantification of the signature in our data and ordering them according to one of the two HRD-related signatures, they show to be in disagreement (Fig 11).

For these reasons, the following step was the *de novo* extraction of signatures directly from HGSC cohort data with additional parameters for short segments. 15 signatures have been identified and quantified for all samples showing a cosine similarity slightly higher. When going to the GenomeSpy representation and ordering samples according to the signature SBS58B (Fig15), there seems to be a high concordance between SBS3 and SBS58B. Moreover, looking at the segmentation of the sorted samples, it is evident the difference between the samples with high levels of SBS3 and SBS58B, which are extremely segmented, and the samples with low levels of SBS3 and SBS58B, which are straighter. In general, it seems that SBS58 signatures can better explain segmentation and copy number characteristics, but further proofs need to be provided.

Another important target gene in HGSC is *CDK12*. Cyclin Dependent Kinases (CDKs) are a group of serine/threonine key regulators of many cellular processes. *CDK12* in particular complexes with cyclin K to regulate gene transcription elongation via phosphorylating RNA polymerase II and translation. It also plays a role in RNA splicing, cell cycle progression, cell proliferation, DNA Damage Response (DDR) and maintenance of genomic stability [33].
TCGA study of HGSC revealed it as a tumor suppressor and appeared to be among the ten most recurrently mutated genes of this type of cancer [34]. *CDK12* regulates the transcription of long DNA repair genes like *BRCA1/2*, *ATR* and *ATM*, genes involved in the HRR pathway. As a consequence, alterations in *CDK12* generate a non-functional HRR pathway, endogenous DNA damage, genome instability and sensitivity to DNA damage agents. The phenotype associated with its inactivation is called tandem-duplicator phenotype (TDP), a genomic signature characterized by copy number gains, Focal Tandem Duplications (FTDs). It is important to notice that FTDs are distinct from duplication observed in *BRCA1*-deficient, *cyclin E1*-amplified or other HRD tumors. Indeed, HRD factors were not found in ovarian CDK12-inactivated tumor samples, and the tumor gene expression profiles have been shown to be different [35].
CDK12-associated FTDs can result in expressed gene fusions and fusion-induced neoantigens, raising the possibility of *CDK12* loss of function alteration as a predictive biomarker for immune checkpoint inhibitor sensitivity [33].
This is another example of a feature that is wanted to be caught with the signatures, so that its identification is easier and would consent the stratification of patients according to these mutations. As a consequence, patients with different genomic features could be treated according to them.

Moreover, copy number signatures could shed light on other useful and targetable characteristics still not identified, as well as identify the features already known.

Future steps

Extraction and quantification of de novo signatures are not sufficient for a complete analysis. Other features must be explored for their characterization, in order to associate them to precise phenotypes and mutational mechanisms as it has been done in [19].
One of the first steps will be the analysis of the distribution of each signature along the genome so that it would be possible to understand if one signature is associated with a focal event or to a global effect such as whole genome duplication or chromosomal chromothripsis. Further inspections include testing the association with known driver genes such as *MYC*, *BRCA*, *RB1*.
Validation of the signature is essential and could be done by the application of the same procedure to TCGA data including only HGSC as tumor type.
Once understood and proposed a plausible etiology for identified signatures, samples can be clustered to test if there are differences in the treatment phase or in the genomic segmentation among different clusters.
Moreover, it would be also important to check if some of the de novo signatures are correlated with the PFI, meaning that these signatures would also predict the aggressivity of the cancer present in the patient.
All these analyses would be useful for the ultimate goal, which is the prediction of the platinum-resistance in patients and the identification of possible therapeutic targets for resistant patients. The correlation of some genomic signature with platinum-resistance would give the possibility to predict it in advance in new patients, shifting the treatment strategy from platinum-taxane chemotherapy.

ACKNOWLEDGEMENTS

I'm extremely grateful to my supervisor prof. Enrica Calura for her suggestion of the laboratory in which I developed my thesis project and for her invaluable patience and feedback. Words cannot express my gratitude also for prof. Sampsa Hautaniemi who generously gave me the opportunity to join his research group and take part to the DECIDER project, which also gave me the founds for my experience. Moreover, this endeavor would not have been possible without University of Padua and Erasmus+ Traineeship program for sustaining financially the journey in Finland.
Additionally, I'd like to acknowledge Kari and Jaana, for mentoring me during the project editing and proofreading help, advice and suggestions; but especially for igniting me the passion for systems biology in ovarian cancer.

Special thanks to all professors and lecturers met in the bachelor's and master's degrees because they helped me to build all skills in the biological field and because they stimulated every day my curiosity and interest in all biological fields but more in general in science. I am also grateful to my closest classmates (Anna, Chiara and Linda) and my roommates (Alessia, Chiara, Martina, Giulia and Gessica), who supported me during the semesters and exam sessions, sharing my fears, joy, complaints and successes.

Lastly, I would be remiss in not mentioning my family, especially my parents, for their financial and emotional support and for believing in me and my career every single day of their lives. I'm also grateful to all my friends, close and far, in Italy or Finland, because they gave me best experiences of my life that I will never forget.
The most important person I would like to thank is Marco, because he was close to me every single day and without which I would not have been able to finish this path.

BIBLIOGRAPHY

[1] U. A. Matulonis, A. K. Sood, L. Fallowfield, B. E. Howitt, J. Sehouli, and B. Y. Karlan, "Ovarian cancer," *Nat. Rev. Dis. Primer*, vol. 2, p. 16061, Aug. 2016, doi: 10.1038/nrdp.2016.61.

[2] M.-A. Lisio, L. Fu, A. Goyeneche, Z.-H. Gao, and C. Telleria, "High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints," *Int. J. Mol. Sci.*, vol. 20, no. 4, p. E952, Feb. 2019, doi: 10.3390/ijms20040952.

[3] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Emerging landscape of oncogenic signatures across human cancers," *Nat. Genet.*, vol. 45, no. 10, pp. 1127–1133, Oct. 2013, doi: 10.1038/ng.2762.

[4] N. Y. L. Ngoi and D. S. P. Tan, "The role of homologous recombination deficiency testing in ovarian cancer and its clinical implications: do we need it?," *ESMO Open*, vol. 6, no. 3, p. 100144, Jun. 2021, doi: 10.1016/j.esmoop.2021.100144.

[5] M. A. Khan *et al.*, "Platinum-resistant ovarian cancer: From drug resistance mechanisms to liquid biopsy-based biomarkers for disease management," *Semin. Cancer Biol.*, vol. 77, pp. 99–109, Dec. 2021, doi: 10.1016/j.semcancer.2021.08.005.

[6] A. Davis, A. V. Tinker, and M. Friedlander, "'Platinum resistant' ovarian cancer: What is it, who to treat and how to measure benefit?," *Gynecol. Oncol.*, vol. 133, no. 3, pp. 624–631, Jun. 2014, doi: 10.1016/j.ygyno.2014.02.038.

[7] Y. Li *et al.*, "Patterns of somatic structural variation in human cancer genomes," *Nature*, vol. 578, no. 7793, pp. 112–121, Feb. 2020, doi: 10.1038/s41586-019-1913-9.

[8] S. Lauer and D. Gresham, "An evolving view of copy number variants," *Curr. Genet.*, vol. 65, no. 6, pp. 1287–1295, Dec. 2019, doi: 10.1007/s00294-019-00980-0.

[9] L. Harbers, F. Agostini, M. Nicos, D. Poddighe, M. Bienko, and N. Crosetto, "Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas," *Front. Oncol.*, vol. 11, p. 700568, 2021, doi: 10.3389/fonc.2021.700568.

[10] T. I. Zack *et al.*, "Pan-cancer patterns of somatic copy number alteration," *Nat. Genet.*, vol. 45, no. 10, pp. 1134–1140, Oct. 2013, doi: 10.1038/ng.2760.

[11] S. S. Ho, A. E. Urban, and R. E. Mills, "Structural variation in the sequencing era," *Nat. Rev. Genet.*, vol. 21, no. 3, pp. 171–189, Mar. 2020, doi: 10.1038/s41576-019-0180-9.

[12] D. A. Peiffer *et al.*, "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping," *Genome Res.*, vol. 16, no. 9, pp. 1136–1148, Sep. 2006, doi: 10.1101/gr.5402306.

[13] D. M. Bickhart, Ed., *Copy Number Variants: Methods and Protocols*, vol. 1833. New York, NY: Springer, 2018. doi: 10.1007/978-1-4939-8666-8.

[14] L. B. Alexandrov *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, no. 7793, pp. 94–101, Feb. 2020, doi: 10.1038/s41586-020-1943-3.

[15] C. D. Steele, N. Pillay, and L. B. Alexandrov, "An overview of mutational and copy number signatures in human cancer," *J. Pathol.*, vol. 257, no. 4, pp. 454–465, Jul. 2022, doi: 10.1002/path.5912.
[16] S. A. Forbes *et al.*, "The Catalogue of Somatic Mutations in Cancer (COSMIC)," *Curr. Protoc. Hum. Genet.*, vol. Chapter 10, p. Unit 10.11, Apr. 2008, doi: 10.1002/0471142905.hg1011s57.
[17] G. Macintyre *et al.*, "Copy number signatures and mutational processes in ovarian carcinoma," *Nat. Genet.*, vol. 50, no. 9, pp. 1262–1270, Sep. 2018, doi: 10.1038/s41588-018-0179-8.
[18] R. M. Drews *et al.*, "A pan-cancer compendium of chromosomal instability," *Nature*, vol. 606, no. 7916, pp. 976–983, Jun. 2022, doi: 10.1038/s41586-022-04789-9.
[19] C. D. Steele *et al.*, "Signatures of copy number alterations in human cancer," *Nature*, vol. 606, no. 7916, pp. 984–991, Jun. 2022, doi: 10.1038/s41586-022-04738-6.
[20] E. Oh *et al.*, "Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples," *PloS One*, vol. 10, no. 12, p. e0144162, 2015, doi: 10.1371/journal.pone.0144162.
[21] A. McKenna *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, doi: 10.1101/gr.107524.110.
[22] H. M. Amemiya, A. Kundaje, and A. P. Boyle, "The ENCODE Blacklist: Identification of Problematic Regions of the Genome," *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Jun. 2019, doi: 10.1038/s41598-019-45839-z.
[23] P. Van Loo *et al.*, "Allele-specific copy number analysis of tumors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 39, pp. 16910–16915, Sep. 2010, doi: 10.1073/pnas.1009843107.
[24] D. L. Cameron *et al.*, "GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly," *Genome Res.*, vol. 27, no. 12, pp. 2050–2060, Dec. 2017, doi: 10.1101/gr.222109.117.
[25] "HMF Tools." Hartwig Medical Foundation, Aug. 05, 2022. Accessed: Aug. 24, 2022. [Online]. Available: https://github.com/hartwigmedical/hmftools
[26] E. N. Bergstrom *et al.*, "SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events," *BMC Genomics*, vol. 20, no. 1, p. 685, Aug. 2019, doi: 10.1186/s12864-019-6041-2.
[27] S. M. A. Islam *et al.*, "Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor." bioRxiv, p. 2020.12.13.422570, Apr. 10, 2022. doi: 10.1101/2020.12.13.422570.
[28] A. R. Shakoori, "Fluorescence In Situ Hybridization (FISH) and Its Applications," *Chromosome Struct. Aberrations*, pp. 343–367, Feb. 2017, doi: 10.1007/978-81-322-3673-3_16.
[29] A. A. Ahmed *et al.*, "Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary," *J. Pathol.*, vol. 221, no. 1, pp. 49–56, May 2010, doi: 10.1002/path.2696.
[30] S. Zaccaria and B. J. Raphael, "Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor

sequencing data," *Nat. Commun.*, vol. 11, no. 1, p. 4301, Sep. 2020, doi: 10.1038/s41467-020-17967-y.
[31] J. Schütte, J. Reusch, C. Khandanpour, and C. Eisfeld, "Structural Variants as a Basis for Targeted Therapies in Hematological Malignancies," *Front. Oncol.*, vol. 9, p. 839, 2019, doi: 10.3389/fonc.2019.00839.
[32] R. R. da Cunha Colombo Bonadio, R. N. Fogace, V. C. Miranda, and M. D. P. E. Diz, "Homologous recombination deficiency in ovarian cancer: a review of its epidemiology and management," *Clin. Sao Paulo Braz.*, vol. 73, no. suppl 1, p. e450s, Aug. 2018, doi: 10.6061/clinics/2018/e450s.
[33] E. S. Sokol *et al.*, "Pan-Cancer Analysis of CDK12 Loss-of-Function Alterations and Their Association with the Focal Tandem-Duplicator Phenotype," *The Oncologist*, vol. 24, no. 12, pp. 1526–1533, Dec. 2019, doi: 10.1634/theoncologist.2019-0214.
[34] Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, Jun. 2011, doi: 10.1038/nature10166.
[35] K. Pilarova, J. Herudek, and D. Blazek, "CDK12: cellular functions and therapeutic potential of versatile player in cancer," *NAR Cancer*, vol. 2, no. 1, p. zcaa003, Mar. 2020, doi: 10.1093/narcan/zcaa003.