LEVERAGING TECHNOLOGICAL OPPORTUNITIES IN CANCER GENOMICS

JOSE A. ESPEJO VALLE-INCLAN

LEVERAGING TECHNOLOGICAL OPPORTUNITIES IN CANCER GENOMICS

JOSE ANTONIO ESPEJO VALLE-INCLAN

ISBN: 978-94-6421-266-2

PRINTED BY: Ipskamp Printing || www.ipskampprinting.nl

DESIGN AND LAYOUT: Inês Vilalva || www.inesvilalva.com

The research described in this thesis was supported by a grant of the Gieskes Strijbis Foundation and the Dutch Cancer Society (KWF).

LEVERAGING TECHNOLOGICAL OPPORTUNITIES IN CANCER GENOMICS

Het optimaal benutten van technologische ontwikkelingen voor onderzoek naar het kanker genoom
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
dinsdag 20 april 2021 des middags te 12.45 uur

door

JOSE ANTONIO ESPEJO VALLE-INCLAN

geboren op 4 maart 1992
te Sevilla, Spanje

PROMOTOR: Prof. Dr. E.P.J.G. Cuppen

CO-PROMOTOR: Dr. W. P. Kloosterman

CONTENTS

1

INTRODUCTION

1

THE HUMAN GENOME AND GENOMIC VARIATION

A genome contains all the instructions needed to form and maintain an organism. The human genome is constituted by two times 3.2 billion nucleotides, which can be adenine (A), cytosine (C), thymine (T) or guanine (G). These are stored in long stretches that coil around each other forming a double helix of deoxyribonucleic acid (DNA)[1,2]. The sequential order of these nucleotides establishes the differences between not only species, but also individuals. DNA is generally packed into chromosomes. Humans possess 23 pairs of chromosomes in homeostasis, with one of each pair inherited from each progenitor[3]. Parts of these chromosomes are organized into specific sequences called genes. Genes contain the code that the cellular machinery uses to construct proteins, the biomolecules that perform most of the cellular processes. There are about 20,000 protein-coding genes in the human genome, constituting only 1-2% of the total genome[4]. The function of the vast majority of the remaining genome is not fully understood, although it is clear that a large fraction holds a regulatory function for the expression of the genes[5].

The large majority (~99.5%) of the genomic sequence overlaps between two individuals[3,6]. Most of the remaining alterations are neutral and have no effect, but some modify protein function by affecting genes or regulatory elements. Variants are responsible for the diversity in the population[7,8], but sometimes cause or predispose to diseases such as congenital disorders or cancer[8,9]. Variants can be classified through different criteria. Depending on their origin, they can be germline, when inherited from the progenitors and therefore present in all cells of an organism; or somatic, when acquired during the lifetime of an individual through different processes of DNA damage and DNA repair errors. Germline or somatic variants can affect a single base in the genome (single nucleotide variants, SNVs) or several (multi nucleotide variants, MNVs), including small deletions and insertions (indels) (**Figure 1**). These small variants are the most common variants, although most are harmless polymorphisms[7,8,10]. However, variants can also affect large stretches of nucleotides and are then called structural variants (SVs). SVs consist of the rearrangement of large genomic segments (over 30 base pairs) within or between chromosomes[11,12] (**Figure 1**). There are two main classes of SVs: i) balanced or reciprocal, such as translocations and inversions, where the amount of DNA does not change and ii) unbalanced, such as deletions, duplications and insertions, where a change in the amount of DNA of a genome occurs. This amplification or loss of DNA segments is also called a copy number alteration (CNA). SVs are not as common as SNVs or MNVs. However, due to their larger size, they impact more bases in total in the genome[13,14], they have an increased chance of impact on the homeostasis of the cell[15] and they can also be the cause of human diseases such as neurodevelopmental disorders, autism or cancer by directly affecting key genes and regulatory mechanisms[8,11,12,16,17].

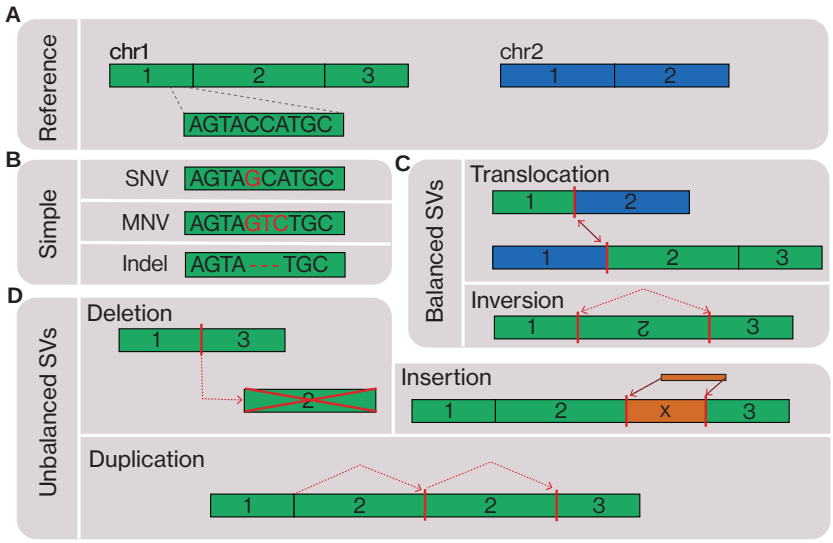


Figure 1: Different types of genomic variation. (**A**) All genomic variation is defined by comparison to an arbitrary reference genome, of which two chromosomes are depicted here. (**B**) Simple variants include single nucleotide variants (SNVs), multi nucleotide variants (MNVs) and short insertions and deletions (indels). (**C**) Balanced structural variants (SVs) do not cause gain or loss of genomic material and include translocations between chromosomes and inversions. (**D**) Unbalanced SVs imply loss or gain of genomic material, and thus entail associated copy number alterations (CNAs). They can be deletions, insertions and duplications.

CANCER IS A GENOMIC DISEASE

Cancer remains a major global health problem with over 18 million cases and 9 million deaths each year[18]. Cancer is a term that englobes different diseases, all defined by an uncontrolled proliferation of transformed cells[9]. Due to a combination of inherited predisposing germline and acquired stochastic somatic SNVs, MNVs, SVs and epigenomic changes, cells may acquire selective growth advantage. Transformed cancer cells also acquire other characteristics to support their abnormal proliferation such as reduced cell death, angiogenesis stimulation, dysregulated metabolism, immune system evasion and migratory characteristics that allow them to metastasize to other tissues[9]. As a result of the uncontrolled proliferation and dysregulated functions, cancer cells often have increased genomic instability and acquire even more somatic mutations. There is then a distinction between passenger events, which are acquired somatic mutations with no effect in tumorigenesis; and driver events, which are mutations that affect key genes and regulatory elements and are able to confer the selective growth advantage[19,20]. Typically there are up to a dozen driver mutations in a tumor, while there can be hundreds to hundreds of thousands passenger mutations

depending on the tumor type[20]. Genes affected by the driver mutations are called cancer driver genes and can be further divided into oncogenes and tumor suppressor genes. Oncogenes acquire activating mutations that confer a positive selection advantage and act in a dominant way. Tumor suppressor genes drive cancer by inactivation or loss of activity often through bi-allelic mutations, which is the case for most familial cancers[20,21]. An example of a cancer driver gene is TP53. This gene encodes for the p53 protein, which monitors genomic integrity in the cell cycle and forces genetically damaged cells into senescence or apoptosis, hence its common moniker: "the guardian of the genome"[22]. The BRAF kinase is also a known cancer driver, prevalent in melanoma and other cancer types[20]. The constitutive activation of BRAF causes a dysregulation of the ERK-MAP signalling pathway, leading to increased cell proliferation and survival. The most common mutation in BRAF is the V600E substitution, although other mutations exist. The mutated protein is specifically targeted with BRAF inhibitors, improving the survival rates of patients with tumors carrying this specific alteration[23]. This example illustrates the importance of the identification of these driver mutations and genes for the treatment and prognosis of cancer, and can only be done through genomic technologies.

During the mid-2000s affordable high-throughput sequencing became available through platforms like 454[24], SOLiD from Thermo Fisher[25], Solexa[26] and several others, encompassed under the umbrella term next-generation sequencing (NGS). Since then, throughput and price of these technologies has steadily improved, with the biotechnological company Illumina (after acquiring Solexa) as market leader. Thus, assays like whole-exome sequencing (WES), where the protein-coding part of the genome is sequenced, and whole-genome sequencing (WGS) have become increasingly important in cancer research. Among countless techniques, knowledge and milestones unlocked by NGS are the development of large international cancer sequencing consortia such as ICGC[27], TCGA[28], PCAWG[19], HMF[29] or several focused on pediatric cancer[30–32]. Through sequencing of these large cohorts our knowledge about the origin, behavior, evolution and actionability of tumors has greatly improved in the last decade. Although genomic data are already being used to guide treatment decisions in a variety of cancers, this is mainly done through targeted gene panels as WES and WGS are less used in clinical care than in research, mainly due to limitations in the availability of fresh-frozen tissue from the tumor biopsies in clinical practice and higher costs when compared to standard diagnostic techniques[33–35]. Nevertheless, WGS is the only technique that fully characterizes the genomic complexity and heterogeneity of a tumor. For this reason, together with technical improvements to reduce costs, there is little doubt that WGS will eventually be implemented routinely in cancer precision medicine[36]. An additional advantage of clinical implementation of WGS is the replacement of many standalone tests for different cancer types or mutation types, bringing in a single test for all cancers and harmonizing protocols for sample collections and data analysis[36].

PRECISION MEDICINE AND LIQUID BIOPSIES

The genomic characterization of tumors allows the development of precision cancer medicine: matching the best drug or treatment with the right patient based on the specific mutations present in the tumor[37–42]. Moreover, the combination of different mutations or the overall mutational landscape of a tumor -also known as mutational signatures[43]- can only be discovered with WGS and also provides therapeutic opportunities[44–46]. Our knowledge of new actionable alterations evolves rapidly, therefore the comprehensiveness of WGS is also essential to be able to reassess data with new knowledge. Particularly, clinical trials can be designed based on particular cancer alterations rather than cancer type to increase sample sizes, which is particularly relevant to rare cancer types[35,47]. Other advantages of WGS implementation are the identification of secondary or multiple alterations to counterpart acquired resistance that emerge after prolonged cancer therapies[35,48] and the application in cancer immunotherapy and antitumor activity of the patients' own immune system through tumor-specific neoantigen prediction[49,50].

Specific molecular profiling of each tumor is key for the development and application of precision oncology. Often, access to tumor material to perform WGS or other analyses is limited and depends on invasive procedures. There is therefore an increasing focus in oncology towards liquid biopsies, which are methods that analyze blood or other biological fluids to derive cancer diagnostic information[51–53]. Due to their low invasiveness, they can be repeated at multiple time points with less inconvenience for the patient. There are different analytes that can be investigated in liquid biopsies: circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), extracellular vesicles or tumor-educated platelets[51]. ctDNA are DNA molecules that are released to the bloodstream by dying cancer cells. These ctDNA fragments reflect comprehensively the genome of the tumor including SNVs, SVs and even other epigenomic alterations[53–55] (**Figure 2A**). Although apoptotic or necrotic normal cells also release their genomic content to the bloodstream (cell-free DNA, cfDNA) ctDNA molecules show a shorter size distribution than normal cfDNA for unclear reasons. The tumor-specific mutations can be detected from ctDNA by sequencing or with other sensitive approaches like digital-droplet PCR. Intra- and inter-tumor heterogeneity and clonality are also reflected in ctDNA[56]. There is a linear relationship between variant allele frequency (VAF) as found in ctDNA and tumor mass or load[57–59]. Furthermore, ctDNA has been found to have prognostic value in clinical trials[60,61]. The main applications of liquid biopsies are early diagnosis of cancer[62,63], treatment selection and monitoring[54,64,65], assessment of prognosis and risk of relapse[66,67] and minimal residual disease and recurrence monitoring through serial measurements[67–69]. The main limitations for further use of ctDNA in the clinic relate to the variation in ctDNA levels and VAF between patients[57,59] and confounding

factors resulting from non-cancer related somatic mutations in different tissues[70,71], leading to suboptimal consistency and precision. With further evidence of clinical utility and multiparametric assays to overcome precision limitations, liquid biopsies will undoubtedly reach their potential in genomic oncology[51,52].

OVARIAN CANCER

Ovarian cancer (OC) is a type of cancer that still has poor survival perspectives. Despite increased knowledge in OC etiology and molecular characteristics, patient survival rates have not improved in the last decades worldwide[72,73]. OC is a very heterogeneous disease with many subtypes. These subtypes differ in their cell of origin, molecular characteristics and disease prognosis[72,74]. Epithelial OC is more common and originates from different cell types surrounding the ovary, mainly from the fallopian tube, while the rarer non-epithelial OC subtypes arise from within the ovary[74,75]. The main histopathological type of epithelial OC is high grade serous carcinomas (HGSCs), which accounts for 70% of OC cases and deaths[73,76]. HGSCs are genomically characterized by somatic TP53 mutations and large copy number aberrations and aneuploidies. Moreover, half of the HGSCs show homologous recombination deficiency, mainly due to mutations in BRCA1 and BRCA2. However, besides these, few genes are recurrently mutated in these HGSCs[77,78]. Furthermore, one fifth of the OC cases are partially explained by germline variants in genes involved in DNA repair like BRCA1, BRCA2, RAD51, PALB1 or CHEK2[79–81]. Other subtypes of epithelial OC are low grade serous carcinomas (LGSC), endometrioid carcinomas (END), clear cell carcinomas (CCC) and mucinous carcinomas (MC). These subtypes harbor somatic mutations in different genes, such ARID1A, PTEN, PIK3CA, CTNNB1 and KRAS[82].

The main treatment for patients with OC consists of debulking surgery and chemotherapy, mainly with a combination of carboplatin and paclitaxel. However, OC often recurs with an acquired chemotherapy resistance, so there is a need for targeted treatments with less harmful side effects. For example several PARP-inhibitors have been recently approved for patients with OC with homologous recombination (HR) deficiencies due to mutations in BRCA1, BRCA2 or other genes in the HR pathway[83–85]. Furthermore, OC, especially HGSCs, presents with high inter- and intra-patient genomic heterogeneity[78,86], highlighting the need of patient-specific treatment approaches to improve prognosis in patients with OC.

There is however limited success in oncology when translating therapeutics and drug development from scientific research to the clinic[87]. Robust model systems of OC,

and other carcinomas, are essential for preclinical cancer biology research and drug development[88,89]. Traditionally, there are two main types of cancer models: cancer cell lines and patient-derived tumor xenografts (**Figure 2B**). Both have contributed enormously to cancer research. Cancer cell lines are derived from tumors and provide fast-growing everlasting models which are relatively easy to handle and experiment with. The first cancer cell line was HeLa, established from a cervical adenocarcinoma biopsy[90]. HeLa cells have been widely used and have contributed to numerous breakthroughs not only in cancer research[91], but also in virology[92] and even in research ethics[93] through a highly controversial trajectory[94,95]. Since, cancer cell lines derived from patient material have been established for different cancer types, including OC[96-99]. Furthermore, cancer cell lines often lose characteristics and cell heterogeneity of the original tumor[100,101]. For example, the two most used HGSC cell lines, SKOV3 and A2780, are not representative of the genomic characteristics of HGSC since they lack TP53 mutations and show near-diploid copy number profiles[100]. Patient-derived tumor xenografts (PDTXs) consist of fresh tumor material implanted subcutaneously or orthotopically into immuno-compromised mice[102,103]. The initial tumor can be then transplanted serially into an increasing number of mice, allowing drug screening in an *in vivo* model that mimics the tumor environment[104,105]. PDTXs have been developed for different OC subtypes, but mainly HGSC[106–108]. They have been shown to mimic chemotherapy and targeted therapy response in HGSC [109–113]. However, the use of PDTXs is expensive and time consuming, making it less suitable for rapid and high-throughput drug screening[114]. Additionally, recent evidence suggests that PDTXs might undergo mouse-specific tumor evolution that differ from that of the patient[115], although that is currently a subject of debate[116].

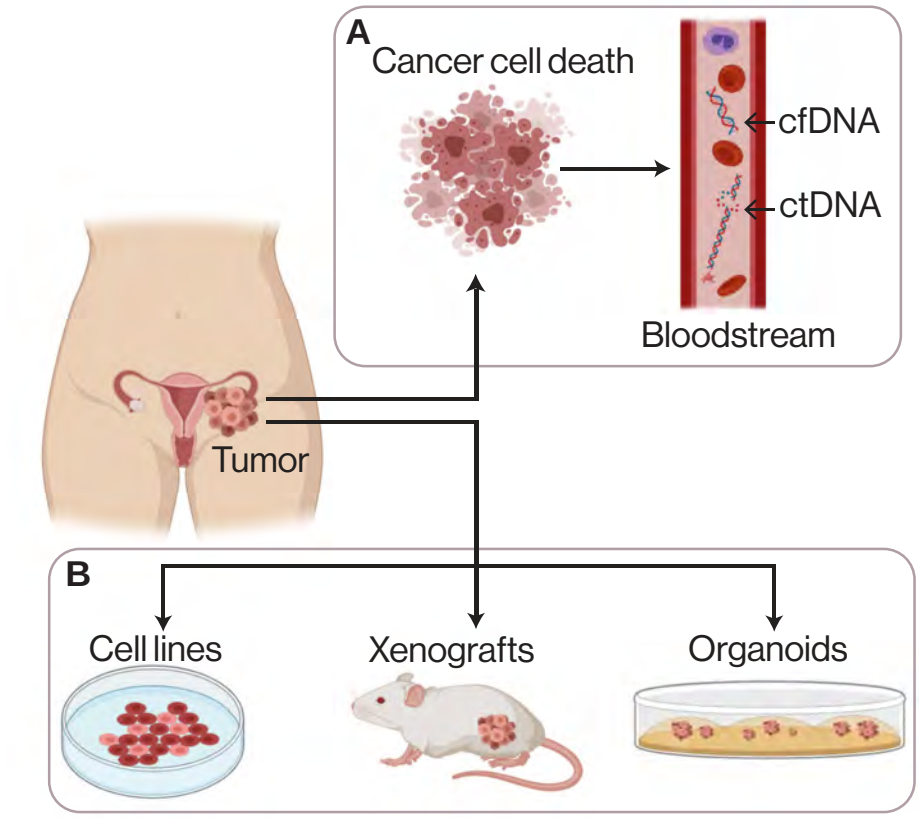


Figure 2: Origin of ctDNA and different cancer models. (A) Dying cells from a tumor, in this case an ovarian tumor, release their fragmented genomic material to the bloodstream. Circulating tumor DNA (ctDNA) carries all the genomic alterations present in the tumor, and can be used as a proxy to estimate tumor presence and dynamics. Cell-free DNA (cfDNA), which originates from other normal cells, is also present in the bloodstream. **(B)** The main patient-derived cancer models include: i) 2D cell cultures that are easy to maintain, manipulate and expand but often do not recapitulate some tumor characteristics, especially after extensive culturing. ii) Xenografts (PDXs), generated by transplantation of tumor cells in immunocompromised mice. They recapitulate tumor characteristics better *in vivo*, but they are costly and low-throughput. iii) Organoids (PDOs), which are 3D cultures of tumor cells that recapitulate faithfully the original tumor and are suitable for high-throughput applications such as drug screening or genetic manipulation. Figure created with BioRender.

ORGANOIDS AS A CANCER MODEL

In the last decade, organoids have emerged as a novel tridimensional model system for normal tissues and cancer (**Figure 2B**). The first organoids were generated from intestinal stem cells, which were grown embedded in Matrigel by mimicking *in vivo* stem cell conditions to form the characteristic crypt-villus structures[117]. Since, organoids have been cultured from a plethora of normal tissues, extensively reviewed in[118], including fallopian tube[119]. Organoids derived from healthy tissue have been shown to expand long-term while remaining genomically stable[120]. They have been used to study mutagenic processes in different tissues[121] and cancer etiology by using genome editing to model colorectal cancer-initiating mutations[122,123].

Organoids can also be generated using patient-derived tumor tissue[124,125]. Successful tumor-derived organoids were first established from patients with colorectal[120], pancreatic[126] and prostate[127] cancer. Furthermore, living tumor-derived (and matched healthy) organoid biobanks have been generated from multiple patients with colorectal[128,129], pancreatic[130–132], liver[133], head and neck[134] and breast[135] cancer. These efforts have shown that tumor-derived organoids maintain the genomic and phenotypical characteristics of the original tumor they were derived from even after long term culturing[124,125,136]. Tumor-derived organoids have also been used for high throughput drug screening and can be xenografted into mice to assess *in vivo* drug response while maintaining the original histopathological characteristics[129].

There are several limitations of organoid technology for (translational) cancer research. Tumor organoid cultures might suffer overgrowth by healthy cells with less mitotic error rates, which can be circumvented with highly pure tumor starting material or culture with selective conditions depending on the driving mutations of the original tumor[120,123,129,135]. Also, organoid cultures are more time- and resource-intensive compared to traditional cell lines, and the tumor microenvironment is still not reflected, with missing blood vessels or immune cells[136], although efforts towards co-cultures that better mimic the *in vivo* situation are underway[137].

Importantly, different tumor-derived organoids have been shown to recapitulate drug responses in the clinic[138–140]. Since patient-derived tumor organoids are genomically and phenotipically closer to original tumors than 2D cell lines, they might be a better model for drug screening while maintaining high-throughput characteristics. Furthermore, if they could be derived from patients in a sufficiently rapid manner, they might be used to study patient-specific tumor characteristics and to rapidly screen for the effectiveness of targeted treatments. Some of the different tumor-derived organoids have been readily

used to explore targeted treatments. Drug screening in breast cancer organoids with several drugs targeting the HER signalling pathway revealed that specific organoid responses depend on HER2 expression status[135]. Separately, prostate cancer organoids with androgen receptor (AR) amplification showed higher sensitivity to enzalutamide, an AR inhibitor, than the rest of the organoids[127]. Overall, despite several limitations and awaiting larger cohort clinical characterizations, organoids have emerged as exciting and promising *in vitro* models to bridge the gap between bench cancer research and clinical care.

STRUCTURAL VARIATION IN CANCER

One of the aforementioned hallmarks of cancer, particularly relevant in OC, is chromosomal instability[9]. Directly related to genomic instability are SVs and in particular somatic SVs. As previously discussed, SVs and copy number alterations (CNAs) have a critical role in tumorigenesis[141] and most tumors bear a considerable amount of somatic SVs. Furthermore, it has been shown that some cancers are primarily driven by somatic SVs, e.g. high grade serous ovarian[77], esophageal[142], neuroblastoma[143], small-cell lung cancer[144] and triple-negative breast cancer[145], with the majority of somatic SV events being non-recurrent even within cancer types[146].

Apart from the simple deletion, inversion, translocation, duplication and insertion SV events previously described, more complex SV phenomena with multiple clustered genomic rearrangement are very common in cancer. Breakage-fusion-bridge (BFB) events, which are the consequence of cycles of DNA breakage and sister-chromatid fusion that cause a dicentric chromosome that bridges during cell division, leading to further breakages. BFB events show oscillating copy number changes surrounded by fold-back inversions[147–150]. Chromoplexy is a series of chained rearrangements involving several chromosomes resulting from simultaneous double stranded breaks that are erroneously repaired[150,151]. Chromothripsis involves a single chromosome shattering which creates up to hundreds of rearrangements in a single event. Chromothripsis is localized to a single or a few chromosomes[152–155], and recent estimates show that up to 40% of all tumors display chromothripsis events[156].

There are diverse mechanisms through which somatic SVs can drive cancer[146]: oncogene amplification via copy number alterations[157]; oncogenic gene fusions through deletions, translocations or inversions[158]; tumor-suppressor inactivation through deletion or gene disruption, often coupled to inactivating mutations in the other allele, common for TP53 or BRCA1 inactivation in ovarian and breast cancer[77,159]; or promoter, enhancer or other

regulatory elements hijacking[160]. Overall genomic instability metrics or specific somatic SV signatures that reflect specific DNA damage can be clinically useful biomarkers[161].

There are examples of therapeutically actionable somatic SVs regardless of their oncogenic mechanisms. For example, the amplification of ERBB2 causes overexpression of HER2 in breast cancer patients, which can be treated with trastuzumab[162]. Also, the prototypical BCR-ABL1 gene fusion in chronic myeloid leukemia, also known as the Philadelphia chromosome, is treatable with the specific inhibitor imatinib[37]. SVs also represent an exciting opportunity in liquid biopsies. Copy number alterations and chromosomal rearrangements can be identified through WGS or targeted approaches in ctDNA from plasma[55,163,164]. Furthermore, somatic SVs may serve as a different type of tumor-specific biomarker to detect and quantify ctDNA with high sensitivity in liquid biopsies through junction-spanning quantitative and highly sensitive PCR assays[165,166].

Traditionally, SVs in cancer have been clinically detected using cytogenetic techniques like fluorescent in situ hybridisation (FISH)[167,168], which works well for hematological malignancies but is restricted to specific known targets and has limited application in solid tumors[146]. An alternative technology is array comparative genomic hybridisation (aCGH), which offers higher throughput and has been applied in both research and the clinic[146,169,170]. However, aCGH is only applicable to unbalanced SVs and suffers from low resolution and higher costs than sequencing approaches[146]. Furthermore, sequencing approaches offer the possibility to identify other alterations than SVs in the same assay. Therefore, most of our knowledge about the role of SVs in cancer genomes stems from the analysis of short-read WGS data, which can be supported by additional sample preparation techniques like paired-end sequencing or long-insert mate pair sequencing to enhance the SV detection power[155,171,172] (**Figure 3**).

There are however some inherent pitfalls to short-read sequencing technologies regarding SV detection. The main problem is that SVs, especially in the germline, are enriched in the vicinity or within repetitive DNA[16,173]. Repetitive DNA complicates short-read data analysis since reads may map to different locations in the genome, with a greater chance of ambiguous mapping with shorter read length[8,174,175](**Figure 3**). Furthermore, the short read-length hinders mapping across SV breakpoints. As a result, there are genomic regions and variation that might have remained inaccessible due to these technological constraints[176–178].

LONG-READ SEQUENCING

In the last several years new sequencing technologies have become available that alleviate these biases, which are often referred to as third-generation or long-read sequencing technologies[177]. One approach is to generate synthetic long reads using long DNA fragments that are sheared and barcoded separately. The barcoded sheared fragments can be then sequenced on existing short-read sequencers. The relation of the short reads can be thus reconstructed by using the shared barcodes (**Figure 3**), hence the name linked reads[179,180]. This technology is commercialized by 10X Genomics and also has important applications for single-cell sequencing[181,182].

The other two main long-read sequencing technologies are from Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT). Both have in common that they do single-molecule sequencing without relying on amplification of DNA fragments and thus eliminating those biases and allowing the sequencing of longer stretches of DNA, routinely reaching average read lengths of 10kb and over 100kb[183,184]. They also share a main limitation, the higher per-base error rate (typically 5-15%) which is also biased in indels and low complexity genomic regions[177]. Also, the higher amount of input DNA required to avoid the amplification step might be a limitation for cancer medicine applications. The approach from PacBio, named single-molecule real-time (SMRT), uses very small wells with a fixed polymerase that adds labelled nucleotides to a single DNA strand, coupled to an imaging system[185,186]. Furthermore, to mitigate the error rates a circular consensus sequencing (CCS) can be generated, where a circularized DNA molecule passes multiple times through the polymerase[184]. Nanopore sequencing records the change of electrical current when a DNA molecule traverses biological pores embedded in a membrane. The distinct voltage change depends on the composition of the specific nucleotide(s) present in the pore at a particular moment[187,188]. Therefore, ONT has unlocked new genomic applications such as direct measurements of RNA molecules[189–191] and RNA and DNA modifications[192–195]. Nanopore sequencing offers improved read lengths, lower costs and higher throughput than PacBio, with similar error rates but a more biased error profile[177]. Efforts to improve the error profiles include pore bioengineering and specific base-calling models.

Long-read sequencing technologies have contributed to expanding the knowledge about structural variation in the last years. Thanks to the increased read length, unambiguous split-mapping across SV break-junctions is easier to achieve (**Figure 3**). Using these new technologies, new areas of the genome previously hidden become accessible like segmental duplications and centromeres [196–198]. Long-read sequencing studies in healthy populations estimate over 25,000 SVs in the germline genome, which is a 3 to 7 fold

increase in SV detection over previous studies using NGS[196,199–203]. Other studies have proven the efficacy of long reads to detect SVs relevant in congenital disease[204–207]. Similarly, recent works have explored the use of long-reads to uncover somatic SVs in cancer genomes[208–211].

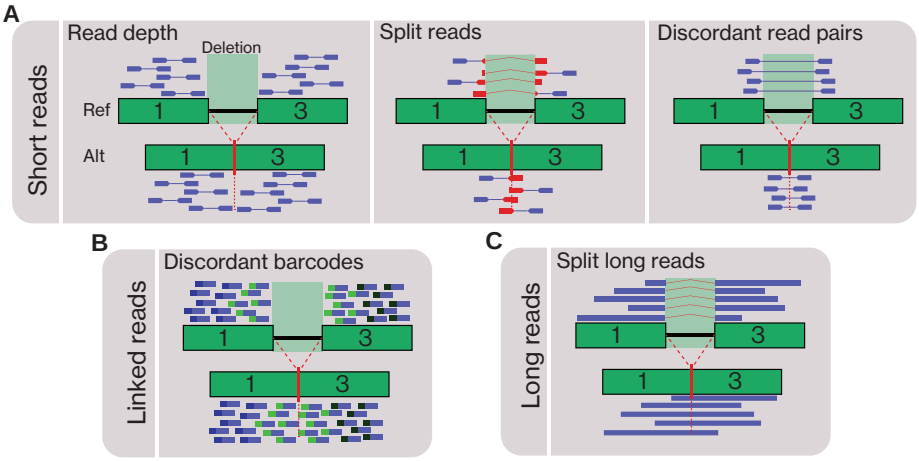


Figure 3: Structural variation detection using different sequencing technologies. Structural variants (SVs) are detected using different approaches for the different sequencing data, as depicted here for a deletion (similar approaches apply to other types of SVs). **(A)** To detect SVs from short-read sequencing data, there are three inference approaches: i) Differences in the sequencing depth, which only works for unbalanced SVs such as deletions (lower sequencing depth), insertions and duplications (increased depth). ii) Split read alignment where a read maps separately in different genomic locations. It is algorithmically challenging to unequivocally map the shorter fragments of a split read. iii) Unexpected distance and orientation between paired reads if using paired-end sequencing **(B)** Inconsistencies within barcode groups distance and orientation are used to detect SVs from linked-reads **(C)** SVs can be detected from long-read sequencing by using the split alignments, similarly to short-read data. Unambiguous mapping of these split reads is less challenging due to the increased length of the fragments.

To unlock the full potential of these long-read sequencing technologies, dedicated bioinformatics pipelines are needed. Although the core algorithms are similar, these new data types present new challenges due to their increased read length and higher error rates[176]. Long-read sequencing has proven to be suitable in a clinical setting[203,212], however analysis pipelines and best practices need to be standardized. In order to validate and benchmark different technologies and analysis pipelines, the availability of high-quality unbiased reference sets is essential[175]. For SVs, germline benchmarks that include orthogonal data from multiple sequencing technologies have become available[199,213]. However in cancer genomics this type of benchmarks are only available for SNVs[214], therefore hindering the advancement of somatic SV discovery and application.

THESIS OUTLINE

This thesis focuses on genomic technology and its application in advancing cancer research and care, presenting several resources and techniques relevant for the community. In Chapter 2, we present an OC organoid biobank and demonstrate that the genomic and transcriptomic characteristics from the tumor of origin are maintained in the corresponding organoids. In Chapter 3, we expand on the application of OC organoids and include drug screening, demonstrating genomic and response heterogeneity in organoids also present in patients. In Chapter 4, we integrate genomic data from multiple technologies to develop a somatic SV truth set essential for the development and benchmarking of somatic SV calling methods. In Chapter 5 we use the nanopore sequencing technology to develop an assay to rapidly detect somatic SVs from a tumor. These somatic SVs were then used to track the dynamics of a tumor. In Chapter 6 we describe the application of targeted nanopore sequencing in pediatric leukemia to close gaps in current diagnostic approaches to detect genomic biomarkers for minimal residual residual disease. Finally, in Chapter 7 I reflect on the results discussed in the rest of chapters and pose challenges and limitations for future research.

2

AN ORGANOID PLATFORM FOR OVARIAN CANCER CAPTURES INTRA- AND INTERPATIENT HETEROGENEITY

An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity

Oded Kopper[1,2], Chris J. de Witte[3*], Kadi Lõhmussaar[1,2*], Jose Espejo Valle-Inclan[3*], Nizar Hami[2,4], Lennart Kester[1,2], Anjali Vanita Balgobind[1,2], Jeroen Korving[1,2], Natalie Proost[5], Harry Begthel[1,2], Lise M. van Wijk[6], Sonia Aristín Revilla[1,2], Rebecca Theeuwsen[5], Marieke van de Ven[5], Markus J. van Roosmalen[3], Bas Ponsioen[2,4], Victor W. H. Ho[7], Benjamin G. Neel[7,8], Tjalling Bosse[9], Katja N. Gaarenstroom[10], Harry Vrieling[6], Maaike P. G. Vreeswijk[6], Paul J. van Diest[11], Petronella O. Witteveen[12], Trudy Jonges[11], Johannes L. Bos[2,4], Alexander van Oudenaarden[1,2], Ronald P. Zweemer[13], Hugo J. G. Snippert[2,4], Wigard P. Kloosterman[3,$] and Hans Clevers[1,2,14,$]

[*]These authors contributed equally to this work
[$]corresponding authors

[1] Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences and UMC Utrecht, the Netherlands
[2] Oncode Institute, the Netherlands
[3] Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, the Netherlands
[4] Molecular Cancer Research, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, the Netherlands
[5] Preclinical Intervention Unit of the Mouse Clinic for Cancer and Ageing (MCCA) at the NKI, Amsterdam, the Netherlands
[6] Department of Human Genetics, Leiden University Medical Center, the Netherlands
[7] Princess Margaret Cancer Center, University Health Network, Canada
[8] Perlmutter Cancer Center, NYU Langone Health, USA
[9] Department of Pathology, Leiden University Medical Center, the Netherlands
[10] Department of Gynecology, Leiden University Medical Center, the Netherlands
[11] Department of Pathology, University Medical Center Utrecht, Utrecht University, the Netherlands
[12] Department of Medical Oncology, Cancer Center, University Medical Center Utrecht, Utrecht University, the Netherlands
[13] Department of Gynaecological Oncology, Cancer Center, University Medical Center Utrecht, Utrecht University, the Netherlands
[14] Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands

Adapted from: Nature Medicine 25:838–849 (2019);
https://doi.org/10.1038/s41591-019-0422-6

ABSTRACT

Ovarian cancer (OC) is a heterogeneous disease usually diagnosed at a late stage. Experimental *in vitro* models that faithfully capture the hallmarks and tumor heterogeneity of OC are limited and hard to establish. We present a protocol that enables efficient derivation and long-term expansion of OC organoids. Utilizing this protocol, we have established 56 organoid lines from 32 patients, representing all main subtypes of OC. OC organoids recapitulate histological and genomic features of the pertinent lesion from which they were derived, illustrating intra- and interpatient heterogeneity, and can be genetically modified. We show that OC organoids can be used for drug-screening assays and capture different tumor subtype responses to the gold standard platinum-based chemotherapy, including acquisition of chemoresistance in recurrent disease. Finally, OC organoids can be xenografted, enabling *in vivo* drug-sensitivity assays. Taken together, this demonstrates their potential application for research and personalized medicine.

INTRODUCTION

Over the past decade, the field of epithelial OC research has gone through a dramatic shift led by a series of recent discoveries[72,73]. It has become clear that OC is a heterogeneous disease consisting of a wide spectrum of distinct molecular and clinical entities. Epithelial ovarian neoplasms can be divided into three main groups: borderline tumors (BTs; non-carcinoma) and type I and type II tumors (carcinomas)[215,216]. BTs account for 15% of OC malignancies and consist primarily of serous BT (SBT) and mucinous BT (MBT) subtypes. BTs are frequently found adjacent to type I tumors and share many of their characteristics. It is therefore believed that they can transform into type I tumors[215]. Type I tumors are genetically stable and carry a distinct set of frequently mutated genes, including, *KRAS, BRAF, PTEN* and *CTNNB1*[215,216]. There are four main type I subtypes: low-grade serous (LGS), mucinous (MC), endometrioid (END) and clear cell (CCC) carcinomas[216]. Type II tumors comprise high-grade serous (HGS) tumors, which are the most common type of OC and account for 70–80% of mortalities[73]. HGS tumors frequently carry mutations in the *TP53* (96%), *BRCA1* and *BRCA2* genes (20%), and are an extreme example of chromosomally unstable cancer[77,217]. HGS tumors are believed to develop either from the fimbria of the fallopian tube (FT)[216] or from the ovarian surface epithelium (OSE). However, the relative contribution of these tissues to tumor development is still under debate[218].

Tumor cell lines and patient-derived tumor xenografts are the most commonly used human model systems for the study of OC[99,219–222]. Despite their contribution to cancer research, these models have a number of drawbacks[223]. Establishing a new cell line is a challenging and time-consuming process that involves a long period of fibroblast contamination reduction and has a low success rate. Thus, in many cases, the resulting cell lines are the product of a strong *in vitro* selection, which inevitably leads to the loss of tumor molecular characteristics, including copy number variations (CNVs), mutations and intrapatient heterogeneity[100]. In contrast to two-dimensional cell lines, xenografts reliably recapitulate components of the tumor environment, such as the three-dimensional structure and the interaction of cancer cells with stroma and blood vessel infiltration[224]. Nevertheless, xenografts involve significant investments in resources for their maintenance, are poorly suited for large-scale drug screening or for genetic manipulation, and undergo rapid mouse-specific tumor evolution[115]. To overcome these drawbacks and to allow personalized approaches to cancer treatment, novel OC research platforms are needed[72,73,224].

As first shown for colorectal cancer[120], tumor organoid cultures represent robust three-dimensional *in vitro* systems that faithfully recapitulate the tumor from which they are

derived[126–129]. Organoid technology is based on the definition of a cocktail of growth factors and small molecules (used in conjunction with the basement membrane mimic Matrigel) to recreate the niche requirements for long-term growth of cells. Organoid cultures can be clonally established from single cells derived from tumor tissue, allowing the study of tumor heterogeneity[125]. Organoids allow rapid assaying of phenotype–genotype correlations and drug sensitivity, while recapitulating patient response[128,135,138,225]. The potential of organoid platforms for OC research was illustrated in a recent paper in which short-term cultured HGS organoids (7–10 d) were genomically characterized and then used in various assays to study DNA repair inhibitor response[226]. Here we present and characterize an OC research platform that supports the efficient derivation and long-term expansion of OC organoids corresponding to non-malignant BTs, as well as MC, CCC, END, LGS and HGS carcinomas.

RESULTS

DERIVATION OF OC ORGANOIDS

OC tissue and blood were obtained from consenting patients who underwent tumor resection and/or drainage of ascites/pleural effusion, either before or after (neoadjuvant) chemotherapy (**Supplementary Table 1**). For each cancer case, the available tissue was used for organoid derivation, DNA isolation and histological analysis. Tumor pieces designated for organoid derivation, were further dissociated and the isolated tumor cells were suspended in basement membrane extract (BME), plated and supplemented with medium (**Extended Data Figure 1a**).

We used a recently described FT organoid medium[119] as our starting point for OC medium optimization. To improve organoid derivation rate, compounds that follow two main guiding criteria were tested as additives to the FT baseline medium: (1) compounds previously reported to be highly expressed in ovarian tumors and therefore hypothesized to support OC growth[227,228] and (2) factors used to support OC cell growth[229,230] and other types of tumor organoids[126,135]. We noted that addition of hydrocortisone, forskolin and heregulinβ-1 to FT medium improved the efficiency of OC organoid derivation. We also observed that Wnt-conditioned medium, an essential component of the FT medium, was not essential for all tumor organoid lines. Moreover, it had a negative effect on some of the lines, presumably due to the presence of serum in the conditioned medium and not Wnt itself. Therefore, we used two types of OC medium for organoid derivation: with ('OCwnt medium') or without ('OC medium') Wnt-conditioned medium (**Supplementary Table 2**). Typically, it became obvious after two to three passages which of the two media was optimal for individual OC cultures. OC organoid growth rates showed

significant variability between cases, with passaging intervals varying from one to four weeks and split ratios ranging from 1:1.5 to 1:4 (**Supplementary Table 3**). Organoids could be expanded long-term, that is, at the time of final submission, 22 lines had been passaged more than 15 times and four lines more than 30 times without slowing down (E**xtended Data Figure 2** and **Supplementary Table 3**). Organoids could be cryopreserved and efficiently recovered (85% success rate, n=33; **Supplementary Table 3**).
OC is often diagnosed after the tumor has already metastasized. In some cases, we were able to obtain both the primary tumor and the different metastatic lesions. We were therefore able to derive multiple organoid lines from individual patients. In one case, we established primary and recurrent tumor organoids from the same patient. In total, we established 56 organoid lines, derived from 32 different patients. Organoids were derived with a success rate of 65%, representing both pre-malignant and malignant neoplasms covering the spectrum of OC, including MBT, SBT, MC, LGS, CCC, END and HGS (**Figure 1a** and **Supplementary Table 4**). OC organoid nomenclature is based on their histopathological subtype and a number that refers to patient and tumor location. Patient clinical data are presented in **Supplementary Table 1**.

DERIVATION OF NORMAL FT AND OSE ORGANOIDS FROM BRCA GERMLINE MUTATION CARRIERS

Women with germline mutations in the *BRCA1*/*BRCA2* genes are at high risk of developing OC[231,232]. Therefore, organoids from normal FT and OSE of these individuals, in addition to non-carriers, should provide a valuable resource for research on the early stages of tumor development. We obtained FT and ovarian tissue from women undergoing prophylactic bilateral salpingo-oophorectomy (pBSO). As previously reported for FT organoids[119], pBSO-derived FT organoids were visible within 3–4 d after isolation, displayed a rounded, cystic phenotype and could be maintained long-term. Consistent with their tissue of origin, FT organoids expressed markers of both secretory and ciliated cells (PAX8 and acetylated-α-tubulin, respectively), and contained beating ciliated cells (**Extended Data Figure 3a–c** and **Supplementary Video 1**). OSE organoids displayed a slower growth rate compared with FT organoids. They were usually visible 1–2 weeks after plating and could be passaged once every 2–3 weeks for extended periods of time. OSE organoids displayed a cystic phenotype and expressed cytokeratin 8, demonstrating their epithelial origin (**Extended Data Figure 3d**). In total, we were able to derive (success rate >90%) FT organoids from ten pBSO-patients and OSE organoids from six pBSOpatients. In addition, we derived two FT lines from non-carriers. Normal organoid nomenclature and patient information data for each line are presented in **Supplementary Table 5**.

MORPHOLOGICAL AND HISTOLOGICAL CHARACTERIZATION OF OC ORGANOIDS

Normal FT and OSE organoid lines consistently displayed a cystic morphology with some epithelium folds and invaginations, which appeared on organoid maturation (**Extended Data Figure 3**). In contrast, OC organoids showed wide morphological variation between and within distinct histological subtype groups (**Extended Data Figure 1b,c**). Most BT organoids were cystic, whereas MC, LGS, END and CCC organoids formed denser organoid structures harboring multiple lumens. HGS organoids presented a wide morphological spectrum, varying from cystic to dense with different degrees of circularity and cellular cohesiveness (**Extended Data Figure 1c,d**). Scanning electron microscopy (SEM) revealed that morphological heterogeneity was not restricted to organoid shape, but also occurred at the cellular level (**Extended Data Figure 1c**). Moreover, SEM showed different degrees of cellular organization, as evidenced by cellular cohesiveness and microvilli alignment.

To compare organoids to their corresponding tumor tissue, we performed hematoxylin and eosin (H&E) staining and evaluated expression of OC protein biomarkers, such as paired box gene 8 (PAX8) and tumor protein p53. Of note, the tumor organoids consist of the transformed epithelial cells of a tumor, but do not contain immune, vessel or connective tissue elements. Histological analysis of the primary tumor tissue used for organoid derivation revealed different degrees of normal cell contamination as indicated by H&E and p53 staining (**Extended Data Figure 2c**). This stressed the need for histological analysis of the primary tissue used for organoid derivation, as low tumor purity can influence organoid derivation efficiency and genomic correlation between organoids and tissue.

H&E staining of OC organoids revealed multiple tumor characteristics, such as the presence of papillary-like structures, nuclear and cellular atypia, and features of hobnail cells (**Figure 1** and **Extended Data Figure 1d**). These characteristics were not detected in normal FT and OSE organoids, which, in contrast, displayed well-organized epithelium (**Extended Data Figure 3**). Moreover, in an H&E-based blinded test conducted by a certified pathologist on samples from normal FT and OSE organoids (n=5) as well as OC organoids (n=18), only FT and OSE organoids were classified as 'normal'. OC organoids were either classified as 'non-definitive' (n=5, 28%) or malignant (n=13, 72%). OC organoids that were classified as 'non-definitive' corresponded to BT and LGS tumors (n=4 and n=1, respectively). In agreement with their histological classification, most MBT and MC organoid lines were positive for periodic acid–Schiff (9 out of 11) and negative for PAX8 (7 out of 11) staining, the latter a hallmark that distinguishes ovarian mucinous and serous tumors (**Figure 1c** and S**upplementary Table 6**)[233]. Ovarian serous organoids that

were tested retained PAX8 and p53 expression status as observed for their corresponding tumor tissue (**Figure 1d,e, Extended Data Figure 2e** and **Supplementary Table 6**). Mutations in the TP53 gene can lead to diverse patterns of p53 staining, such as protein loss or strong nuclear staining. Such patterns were observed in different HGS organoid lines and their corresponding tumor tissue and were in agreement with their sequencing data (**Figure 1e** and **Supplementary Table 7**). Organoids displayed a high percentage of Ki67-positive cells (**Extended Data Figure 2b**). Thus, histological analysis of OC organoids demonstrated their similarity to the carcinoma fields within the corresponding primary tumors and their distinction from non-malignant FT and OSE organoids.

ORGANOIDS FAITHFULLY RECAPITULATE OC AT THE GENOMIC LEVEL

To further validate that OC organoids are composed of malignant cells, we performed metaphase spread analysis. The majority of tested organoid lines were aneuploid, a well-characterized hallmark of most solid tumors[234]. Interestingly, in some cases, a significant variation in average chromosome number was observed for different organoid lines derived from the same patient (**Figure 2a**).

To determine whether OC organoids faithfully recapitulate the genomic landscape of the primary tumors from which they were derived, we next performed whole-genome sequencing (WGS) analysis. In total, we sequenced 40 organoid lines from 22 different patients. The corresponding tumor and normal blood samples for 35 of these lines were also sequenced and used as a reference (**Supplementary Table 7**). We first used WGS data to estimate the percentage of malignant cells in both organoid and tumor samples[29]. As predicted from histological analysis, in most cases, cancer cell content of organoids was considerably higher than that of the corresponding tumor (tumor organoids 88.1 ± 23% versus tumor tissue 45.1 ± 9.2% (mean ± s.d.) across all samples; **Extended Data Figure 2d** and **Supplementary Table 7**). CNV analysis revealed similar patterns between organoid/tumor pairs (**Figure 2b** and **Extended Data Figure 4a**). Moreover, comparing the genomic landscape from early and late passage HGS organoids revealed that CNVs were well maintained even after prolonged passaging (HGS-1, passage eight versus 32; HGS-2, passage six versus 15; HGS-3.1, passage four versus 32; HGS-3.2, passage four versus 25; HGS-6, passage eight versus 21; HGS-1-R2, passage four versus 17; **Figure 2c** and **Extended Data Figure 4a**). Most organoids derived from HGS tumors displayed many CNVs, whereas organoids derived from type I tumors and BTs revealed a relatively subtle number of CNVs (**Figure 2b** and **Extended Data 4a**). Thus, OC organoids recapitulate the genomic characteristics of the different OC subtypes from which they are derived[216,235].

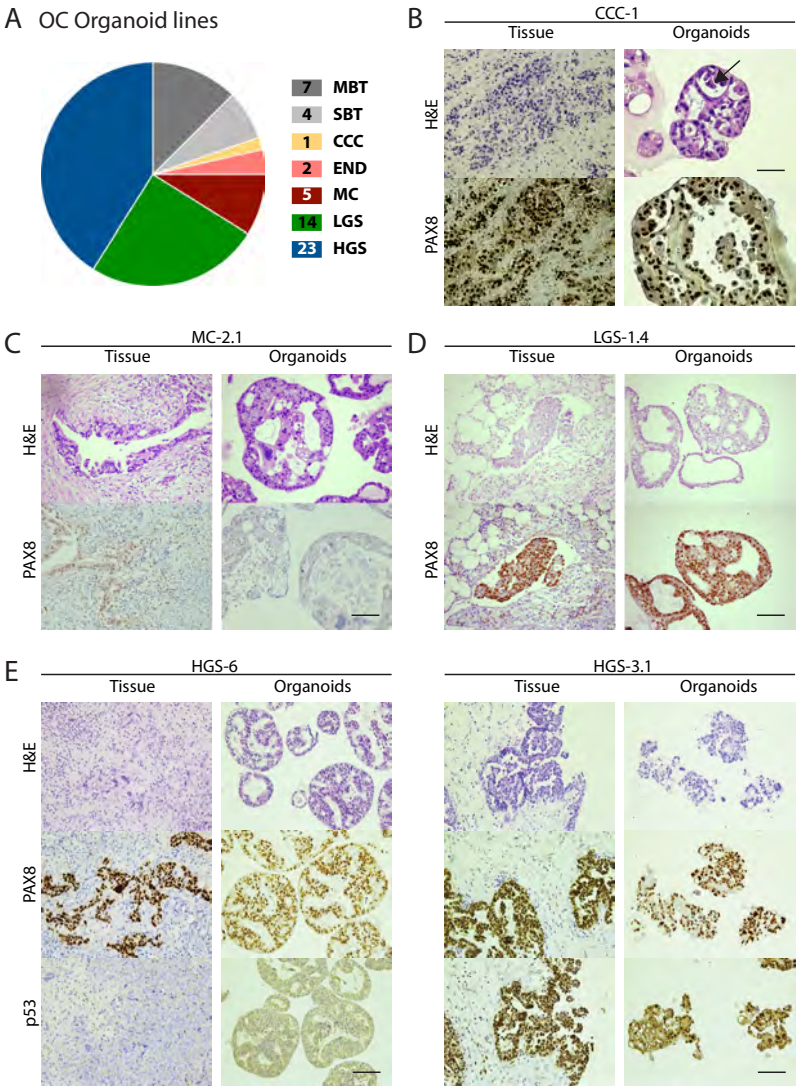


Figure 1: Subtype diversity and histological characterization of OC organoids. **(A)** An overview of established OC organoid lines according to their subtype distribution. Numbers in the legend represent the number of lines established from each subtype. **(B)** Histological comparison of CCC organoids and their corresponding tumor tissue. Top and bottom panels show H&E and PAX8 staining, respectively. Arrow indicates hobnail cells, which characterize CCC. Scale bar, 100 µm. **(C)** Histological comparison of representative MC organoids and their corresponding tumor tissue. Top and bottom panels show H&E and PAX8 staining, respectively. Tumor and organoids were negatively stained for PAX8, a marker of the serous subtype. Scale bar, 100 µm. **(D)** Histological comparison of representative LGS organoids and their corresponding tumor tissue. Top and bottom panels show H&E and PAX8 staining, respectively. Organoids maintain positive PAX8 staining. Scale bar, 100 µm. **(E)** Histological comparison of HGS organoids and their corresponding tumors (HGS-6 on the left and HGS-3.1 on the right). H&E staining of the HGS-6 organoid line showed papillary-like structures growing into the lumen, forming a dense phenotype. HGS-3.1 organoids are characterized with disorganized morphology, which is evident by loss of organoid circularity and cellular cohesiveness. PAX8 positively stains both organoids and the tumor cells within the

tissue. Mutations in the *TP53* gene can lead to protein loss, as presented by the HGS-6 organoid/tumor pair, or strong nuclear staining, presented by the HGS-3.1 organoid/tumor pair. Histological characterization across the different organoid lines is presented in Extended Data Figure 2e and Supplementary Table 6. Scale bar, 100 μm.

To further quantify genetic correlation between organoids and corresponding tumors, we analyzed somatic single nucleotide variants (SNVs) and structural variants (SVs). Most SNVs and SVs present in the original tumor were maintained in the organoids derived thereof, and vice versa (**Extended Data Figs. 4b and 5a**). Shared mutations were also maintained after extended passaging (**Extended Data Figure 4b**). Some organoid lines, such as HGS-19, HGS-3.1 and MC-2.1, presented marked differences with their corresponding tumor sample (**Extended Data Figure 5a**). We believe that these differences result from low tumor cell content within the original tumor samples as evident from their low number of SNVs, SVs and the lack of obvious CNVs (**Extended Data Figure 4**).

Next, we tested whether organoids displayed known OC-associated somatic mutations, amplifications and deletions. Somatic mutations in *KRAS* and *BRAF* genes, which are frequently found in MC and LGS tumors[236,237], were identified in the corresponding organoid subtypes (MC-1, MC-2 (*KRAS*), LGS-5 (*BRAF*); **Figure 3** and **Supplementary Table 7**). Moreover, all organoids derived from HGS tumors showed non-silent mutations including missense, stop gain and frameshifts in the *TP53* gene, in some cases accompanied by the loss of the second allele (**Figure 3** and **Supplementary Table 7**). Amplifications of *MYC* and *CCNE1* as well as loss of *RB1*, *PTEN* and *CDKN2A/B* genes (frequent in HGS tumors[77,238]) were observed (**Figure 3**). These oncogenic modifications were mostly conserved between organoids and corresponding tumors (**Figure 3** and **Supplementary Table 7**).

DNA methylation analysis was performed on a subset of organoids at early and late time points, using Illumina Infinium methylationEPIC 850K BeadChip. Clustering of these organoid samples based on the methylation beta-values demonstrated that organoids maintained their epigenetic profile after extended passaging (**Extended Data Figure 5b**), as found previously for colorectal cancer organoids[125].

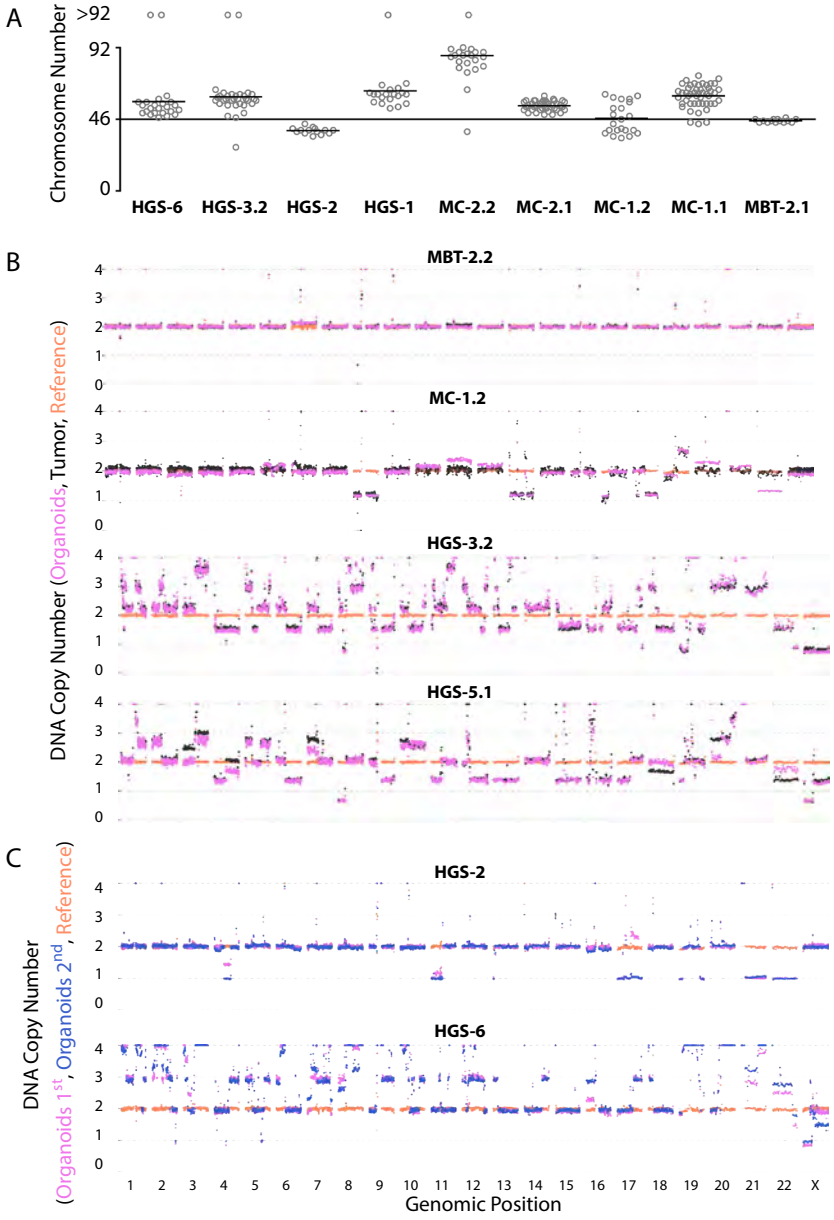


Figure 2: Organoids maintain the genomic landscape of corresponding tumors. (A) Scatter plot presenting chromosome number distribution and mean, based on organoid metaphase spreads. All the lines display aneuploidy except for the BT sample (MBT-2.1). Some of the organoid lines present a relatively narrow chromosome number distribution (MBT-2.1, MC-2.1, HGS-2), whereas others show a wide distribution (MC-1.1, MC-1.2), an indication of tumor heterogeneity. Differences between organoid lines that were derived from a single patient (MC-1.1/MC-1.2 and MC-2.1/MC-2.2) implies intrapatient

heterogeneity. n = number of analyzed metaphase spread, from left to right: 24, 33, 14, 20, 24, 40, 22, 48 and 14. **(B)** Genome-wide CNV analysis of tumor and organoid pairs. For each sample, CNV profile of blood germline reference (orange), tumor (black) and organoids (pink) are displayed. CNVs observed in original tumor samples are maintained in organoid lines. MBT-2.2 organoid line displays a relatively flat CNV pattern in accordance with MBT-2.1 that was derived from the same patient and shows normal metaphase spreads (in Figure 2a). HGS lines display extreme CNV abnormalities (see also Extended Data Figure 4). **(C)** Genome-wide CNV analysis of early (organoids 1st) and late (organoids 2nd) passage organoid pairs (HGS-2, passage six versus passage 15; HGS-6, passage eight versus passage 21). A ploidy of three was assumed for this sample. For each sample, CNV profile of blood germline reference (orange), early (pink) and late (blue) passaged organoid are displayed. CNV profiles observed in organoid samples are maintained.

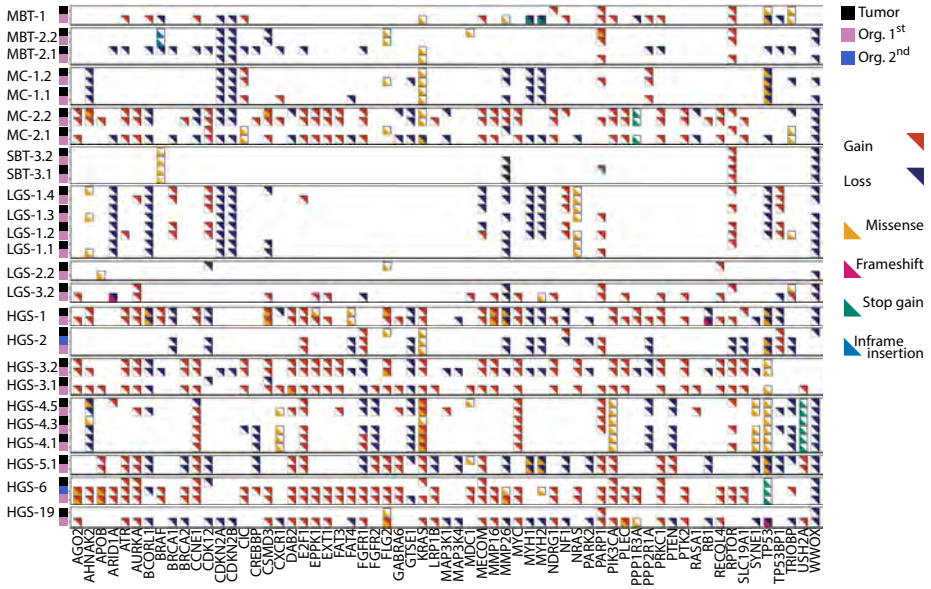


Figure 3: Somatic mutations and amplifications/deletions in OC organoids. Somatic mutations and amplifications/deletions in relevant genes of ovarian cancer. For each sample, tumor/organoid pairs are displayed and indicated by color coding (black, tumors; pink, organoids; blue, organoids re-sequenced and analyzed after extended passaging). Passage number at which organoid lines were sequenced is given in Supplementary Table 7.

OC ORGANOIDS CAPTURE TUMOR HETEROGENEITY

To assess whether organoids capture intrapatient heterogeneity, we compared organoid lines derived from one primary and three metastatic sites of a patient diagnosed with LGS OC (**Figure 4a**). CNV analysis revealed losses and gains shared by all tumor lesions from the same patient (for example, loss of chromosome X) as well as copy number changes only present in the metastatic sites (for example, loss of 17p in LGS-1.2,3,4; **Figure 4a**). These CNVs are conserved between tumor tissue and the

corresponding organoids (**Extended Data Figure 4a**) and, therefore, appear to represent genomic changes that occurred at different time points along the course of tumor evolution. We next tested whether tumor heterogeneity is maintained within an organoid line using a novel single-cell DNA sequencing method (see Methods) and sequenced 791 cells from two recurrent tumor samples (HGS-1-R2 and HGS-1-R3; both were derived from a single patient at different time points) and corresponding organoid lines from either one or two time points (HGS-1-R2, passage five; HGS-1-R3, passage four and 12). Calculation of CNV profiles for each cell was followed by independent component analysis that revealed five distinct clusters (**Figure 4b**).

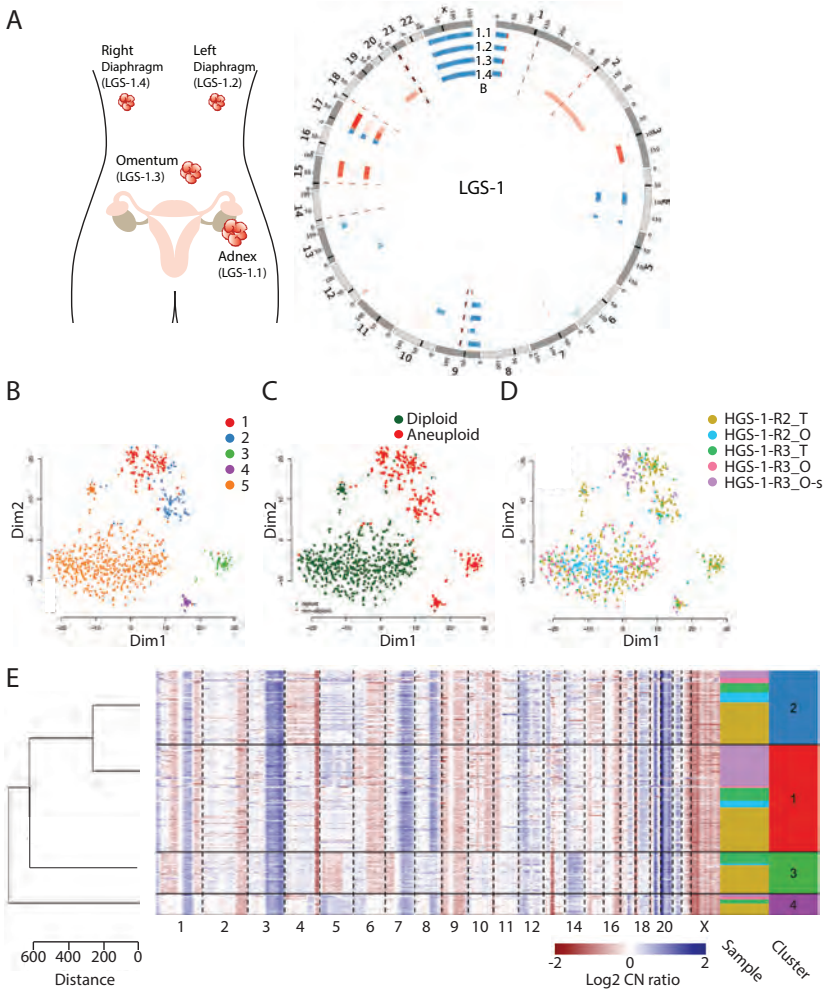


Figure 4: OC organoids capture tumor heterogeneity. (A) Schematic of tumor locations and a circos plot presenting CNV events (red, gain; blue, loss) in the organoid lines derived from a patient diagnosed with LGS OC. Outside to inside: genomic position, LGS-1.1 (adnex tumor), LGS-1.2 (metastasis left diaphragm), LGS-1.3 (metastasis omentum), LGS-1.4 (metastasis right diaphragm), blood germline reference. **(B)** t-SNE

plot of single-cell CNV profiles from two recurrent tumor samples and corresponding organoid lines (HGS-1-R2, HGS-1-R3) of a single patient. Hierarchical clustering has separated the cells into five different clusters (color coded). Total number of analyzed cells is 791. **(C)** t-SNE plot presenting diploid (green) and aneuploid (red) cells. Total number of analyzed cells is 791. **(D)** Single-cell distribution into the different clusters according to sample of origin. T, tumor; O, organoid; -s, second time point analysis. HGS-1-R2_T, n = 351 cells; HGS-1-R2_O, passage 5, n = 159 cells; HGS-1-R3_T, n = 93 cells; HGS-1-R3_O, passage 4, n = 122 cells; HGS-1-R3_O, passage 12, n = 66 cells. **(E)** Clustered CNV heat map of aneuploid cells presenting gains (blue) and losses (red) across the genome. Sample origin and cluster belonging of each cell is color coded.

Clusters 1–4 comprised aneuploid cells whereas cluster 5 comprised diploid cells (**Figure 4c**). As expected, tumor samples that were obtained from ascites drainage of a single patient within a one-month interval overlapped with each other and did not form separate clusters (**Figure 4d**), thus validating the robustness of the single-cell DNA sequencing method. Organoid-derived cells overlapped with the same five clusters (albeit with low representation in cluster 3) demonstrating both their heterogeneity and resemblance to the original tumor samples (**Figure 4d**). HGS-1-R3 relative cell abundance in cluster 5 (diploid cells) was dramatically reduced after extended passaging (passage four versus 12), whereas representation of clusters 1, 2 and 4 (aneuploid cells) increased (**Figure 4d,e**), suggesting that tumor cells overgrew normal cells over time, while maintaining tumor heterogeneity.

GENE EXPRESSION ANALYSIS OF OC ORGANOIDS

To assess organoid gene expression profiles, we performed RNA sequencing (RNA-seq) on 35 OC organoids, and six normal OSE and FT organoids. Hierarchical clustering assigned organoids to three independent main groups, representing (1) HGS carcinomas, (2) MC and END tumors and (3) mainly LGS carcinomas, FT and OSE (**Figure 5**). Organoids derived from multiple tumor lesions of the same patient were transcriptionally more similar to each other than to unrelated organoid lines (for example, MC-1.1,2 and HGS-3.1,2). In a similar manner, organoids that were sequenced at a second time point after extended passaging clustered with their corresponding samples (HGS-1, passage eight versus 32; HGS-3.1, passage four versus 32; HGS-1-R2 passage four versus 17). Of note, non-malignant MBT and malignant MC organoids clustered together. This was seen in eight organoid lines derived from four different patients (two MC and two MBT), suggesting a biological link between these samples. This finding is in agreement with a causality hypothesis that suggests a stepwise progression from BTs to invasive carcinomas[239–241]. Furthermore, OSE(P)7 organoids (derived from a sample collected during risk-reducing salpingo-oophorectomy) clustered together with OC organoids and apart from normal OSE and FT organoid lines. This finding, together with morphological, histological and metaphase spread analysis (**Extended Data Figure 3e,f**), suggested that OSE(P)7 consists of malignant cells that were not diagnosed by routine pathological ex-

amination.

GENETIC MANIPULATION AND DRUG SCREENING OF OC ORGANOIDS

To demonstrate the experimental potential of OC organoids, we next adapted genetic manipulation techniques and drug-screening methods for normal FT and OC organoids. Normal FT organoids were electroporated with pSpCas9(BB)- 2A-GFP plasmid into which we cloned a guide RNA targeting the *TP53* gene (**Extended Data Figure 6a**). Thus, we could determine the electroporation efficiency by monitoring GFP expression (**Extended Data Figure 6c,d**) and target the *TP53* gene, which is believed to be mutated at an early time point in the course of HGS tumor development. Three days after electroporation, nutlin3a (which inhibits MDM2–p53 interaction[242] and, therefore, kills *TP53* wild-type clones) was added to the medium (**Extended Data Figure 6a,b**). Surviving clones were picked, clonally expanded and analyzed for *TP53* mutations (**Extended Data Figure 6e**). As a result, multiple clones harboring mutations in *TP53* from carriers of BRCA germline mutations were established (**Extended Data Figure 6f**). In a similar manner, we have electroporated FT organoids with plasmids targeting both *TP53* and *RB1* genes and established clones in which both genes were knocked out (**Extended Data Figure 6f**). Clone expansion was accompanied by morphological alterations including transition from cystic to denser organoids and increased cell shedding into the organoid lumen (**Extended Data Figure 6g**). Hierarchical clustering based on RNA-seq assigned the clones into different clusters according to their genetic modifications (**Extended Data Figure 6h**).

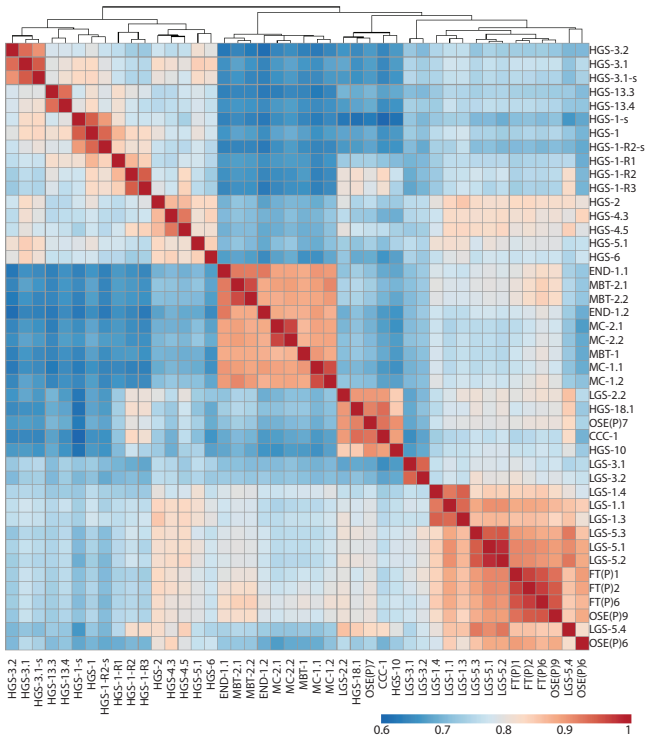


Figure 5: Gene expression analysis of OC organoids. Heat map of Spearman correlation values of normal FT (n = 3 independent FT lines), OSE (n = 3 independent OSE lines), non-malignant BTs (n = 3 independent MBT lines) and malignant organoid lines (n = 32 independent malignant lines), based on RNA-seq expression data. Read counts were normalized for sequencing depth and the 5,000 most-variable genes were used. For three organoid lines, a second time point was analyzed after extended passaging, demonstrating high correlation with early passaged organoids. -s, second time point analysis. HGS-1: passage eight and 32; HGS-3.2: passage four and 32; HGS-1-R2: passage four and 17. Passage number in which all organoid lines were sequenced is given in Supplementary Table 7.

To demonstrate that OC organoids can be genetically modified in a stable manner, they were transduced with a lentiviral vector driving expression of fluorescently tagged histone-2B (H2B-Neon). H2B-Neon-transduced organoids enabled three-dimensional live cell imaging of mitosis and revealed multiple aberrant chromosomal segregation events (Supplementary Videos 2–6).

Next, we tested organoid sensitivity to platinum/taxane drugs that are commonly used in OC treatment protocols, that is carboplatin, paclitaxel, as well as non-platinum/taxane drugs that previously have been suggested as possible treatments for OC. The drug panel included drugs targeting the PI3K/AKT/mTOR pathway (alpelisib, pictilisib, MK2206, AZD8055), poly (ADPribose) polymerase (PARP) (Niraparib), the tyrosine kinase Wee1 (adavosertib) and gemcitabine. Organoids were disrupted into small clumps and dis-

pensed into 384-well plates pre-coated with BME. A cell viability assay was performed 5 days after the drugs were added and organoid drug sensitivity was represented by the average area under the dose–response curve (AUC) of two technical replicates[243]. Assay quality was confirmed by calculating plate Z-factor across all plates (mean=0.61; **Extended Data Figure 5e**) and by the correlation of AUC between technical and biological replicates (Pearson correlation=0.94, 0.87, respectively; **Extended Data Figure 5c,d**).

Unsupervised hierarchical clustering based on platinum/taxane drug sensitivity divided the organoids into two main clusters: sensitive lines that consisted primarily of HGS organoids and resistant lines that consisted primarily of non-HGS organoids (**Figure 6b**). Notably, the HGS-1-R3 line, which was derived from ascites of recurrent disease, clinically resistant to chemotherapy (**Supplementary Table 1**), clustered together with the resistant cluster. HGS-1 line, which was derived from the primary, chemotherapy- sensitive tumor of the same patient clustered with the sensitive cluster (**Figure 6a,b**).

Since the *TP53* gene is mutated in the vast majority of OC, we tested whether nutlin3a can serve to rapidly distinguish between wild-type and mutated *TP53* OC organoids. In total, 16 organoid lines were tested (3 normal FT lines, one genetically modified FT clone and 13 OC lines). As expected, all FT organoid lines were highly sensitive to nutlin3a treatment whereas the genetically modified clone in which we knocked out the *TP53* gene and the OC lines (with one exception) were resistant (**Figure 6c,d**). The only OC line that was sensitive to nutlin3a, was LGS-1.3 and in this organoid, indeed no point mutation in the *TP53* gene was identified (**Supplementary Table 7**).
Drug-screening assays demonstrated differential drug responses of individual organoid lines (**Figure 6a–e**). For example, HGS-3.1 organoid line was highly sensitive to gemcitabine, adavosertib, carboplatin and paclitaxel and resistant to drugs that target the PI3K/AKT/mTOR pathway, whereas HGS-23 line demonstrated the opposite drug sensitivity pattern (**Figure 6a–d**).

Homologous recombination-deficient cells have been shown to be sensitive to PARP inhibitors[244,245]. To determine whether this correlation is also present in OC organoids, a subset of organoid lines with differential responses to niraparib (**Figure 6e**) was tested for homologous recombination by using the recombination capacity (RECAP) test, which assesses homologous recombination capacity using accumulation of RAD51 protein at sites of DNA double-strand breaks[246]. Organoids were irradiated with 5 Gy X-rays, recovered for two hours, fixed and stained with antibodies against RAD51 and geminin (a marker for S/G2 phases of the cell cycle). The percentage of geminin+ cells with RAD51 foci was scored blinded for sensitivity to niraparib. Organoids with a low percentage of geminin+ cells with RAD51 foci were more sensitive to niraparib com-

pared with organoids with a high percentage of geminin+ cells with RAD51 foci (with the exception of MC-2.1) (**Figure 6e**).

XENOTRANSPLANTATION OF OC ORGANOIDS AND IN VIVO DRUG SENSITIVITY

We next tested whether OC organoids can be orthotopically or subcutaneously transplanted into immunodeficient mice. For orthotopic transplantations, organoids were transduced with a lentiviral vector encoding luciferase and transplanted into the mouse bursa. Bioluminescence imaging was used to validate tumor growth (**Extended Data Figure 5f**). All three lines that were orthotopically transplanted grew into a tumor (**Supplementary Table 8**). Six out of seven lines were successfully transplanted subcutaneously (**Supplementary Table 8**). Histological analysis of orthotopically transplanted HGS carcinoma organoid line demonstrated that the tumor invaded the ovary, displayed prominent nuclear atypia, slit-like spaces and maintained PAX8 and p53 staining (**Figure 6f** and **Extended Data 5g**). The MC organoid line that was subcutaneously transplanted showed characteristics of a MC tumor including goblet cells and haphazardly arranged neoplastic glands lined by columnar cells (**Extended Data Figure 5h**).

To validate whether *in vitro* drug sensitivity is recapitulated *in vivo*, we chose the HGS-3.1 organoid line that was highly sensitive to gemcitabine (**Figure 6c**), a nucleoside analog that is in clinical use for HGS OC. Organoids were subcutaneously injected and tumor size was monitored. Once it reached 50 mm3, mice were randomly selected and treated with vehicle or gemcitabine. While tumors continued growing in vehicle-treated mice, tumor growth was completely blocked or reduced in gemcitabine-treated mice, as indicated by tumor size measured at the end of the experiment (vehicle and gemcitabine-treated mice, n=9 and n=7, respectively) (**Figure 6g**).

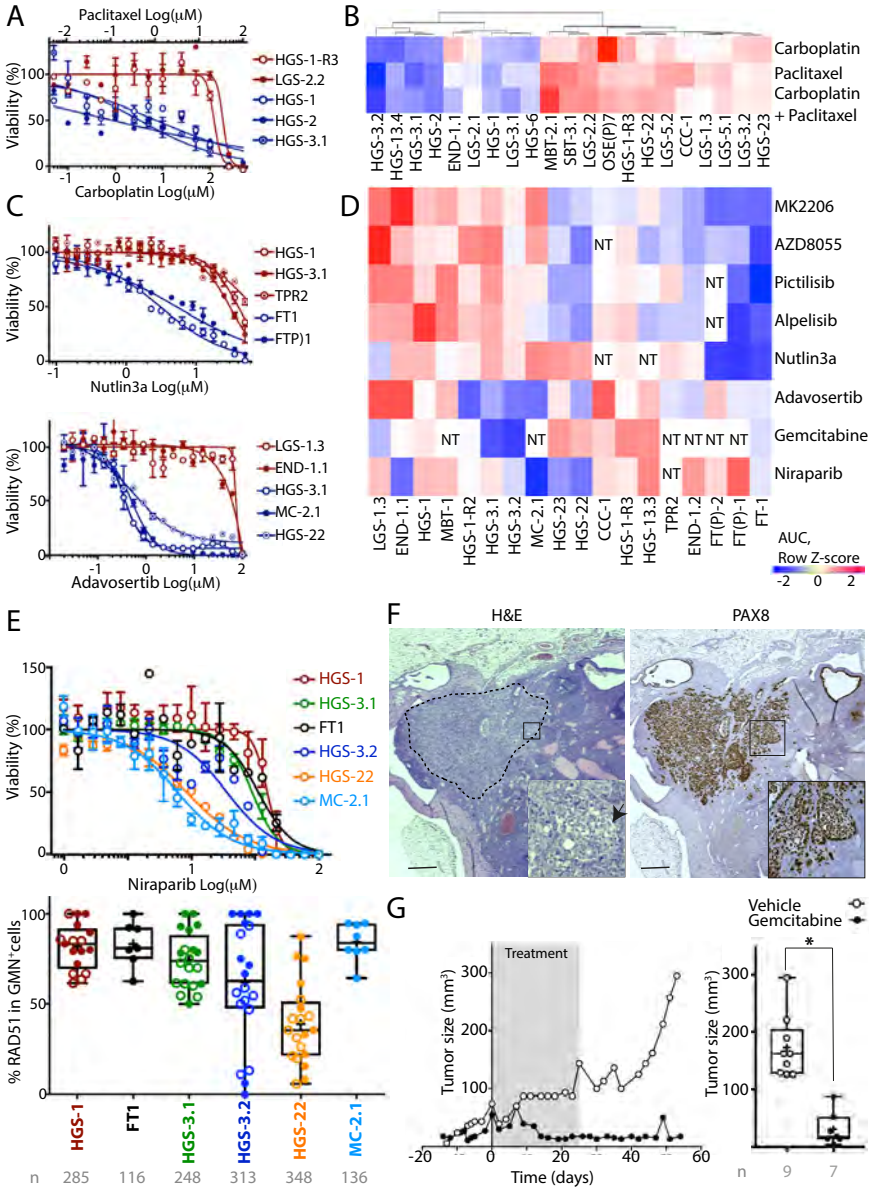


Figure 6: *In vitro* and *in vivo* drug sensitivity assays. **(A)** Representative dose–response curves of HGS and LGS organoid lines treated with carboplatin/paclitaxel. Organoid line derived from a recurrent disease (HGS-1-R3) show acquired resistance. Dots represent the mean of technical duplicates. Error bars represent s.e.m. of technical duplicates. **(B)** Heat map of Euclidean distance of 21 distinct organoid lines, based on AUC row Z-score values. As expected, most HGS OC organoids (6 out of 9) are more sensitive to carboplatin/paclitaxel drugs compared with non-HGS OC organoids (9 out of 12). The HGS-1 organoid line is sensitive to carboplatin/paclitaxel drugs, whereas the matching recurrent organoid line (HGS-1-R3) is resistant. **(C)** Representative dose–response curves for nutlin3a (top) and adavosertib (bottom). Normal FT organoids show high sensitivity for nutlin3a, whereas HGS and genetically modified FT line, which are

mutated in the TP53 gene, are resistant. Dots represent the mean of technical duplicates. Error bars represent s.e.m. of technical duplicates. **(D)** Heat map of Euclidean distance, based on AUC row Z-score values, showing organoid response to a panel of drugs, including PI3K/AKT/mTOR pathway, PARP and Wee1 inhibitors. n = 18 distinct organoid lines. NT, not tested. **(E)** Top, dose–response curves for niraparib show differential response between organoid lines. Dots represent the mean of technical duplicates. Error bars represent s.e.m. of technical duplicates. Bottom, box-and-whisker plot (minimum to maximum) presenting RAD51 foci score after radiation. Each point represents percentage of RAD51+ cells within geminin+ (GMN+) cell population in one organoid. Horizontal bars and '+' represent median and mean of all dots, respectively. Empty and full dots show results of two biologically independent experiments conducted one or two passages apart. Total number (n) of analyzed geminin+ cells in each organoid line is presented. **(F)** Histological analysis of organoid-derived xenograft (HGS-3.1) following orthotopic transplantation into the mouse bursa. Tumor cells have invaded into the mouse ovary and H&E staining (left) shows solid pattern with indications for slit-like spaces (arrow) as well as pleomorphic cells with prominent nuclear atypia. Xenograft has maintained PAX8-positive staining (right). A summary of organoid-derived xenograft experiments is presented in Supplementary Table 8. Scale bar, 0.5 mm. **(G)** Gemcitabine-sensitive organoids were subcutaneously injected into immunodeficient mice and tumor size was monitored. Once the tumor reached 50 mm3, mice were randomly selected and treated with intraperitoneal injections of gemcitabine (2 mg per kg body weight) (n = 7 independent mice) or vehicle (n = 9 independent mice), 5 times per week for 4 consecutive weeks (in total 20 injections). Left, an example of tumor growth over time in a vehicle (white dots) and a gemcitabine-treated (black dots) mouse. Right, box-and-whisker plot (minimum to maximum) summarizing the results across all vehicle and gemcitabine-treated mice, showing tumor size at day 55. Horizontal bars and '+' represent the median and mean of all dots, respectively. *P < 0.001, t-test.

DISCUSSION

Developing reliable experimental models that address clinical challenges, such as early detection, tumor recurrence and acquired chemotherapy resistance, is a high priority in OC research[73]. In this study, we describe an organoid platform that enables long-term *in vitro* expansion, manipulation and analysis of a wide variety of OC subtypes. A comprehensive analysis demonstrates that OC organoids maintain tumor histological characteristics, such as nuclear and cellular atypia, and biomarker expression, such as p53 and PAX8. Organoids and corresponding tumors remained highly similar at the genomic level, even after extended passaging. Furthermore, organoids recapitulated OC hallmarks, such as CNVs, recurrent mutations and tumor heterogeneity. Finally, unsupervised hierarchical clustering of gene expression data grouped the organoids according to their tumor type and demonstrated that LGS organoids are more similar to normal samples than are HGS lines.

During organoid biobanking of normal FT and OSE samples, obtained from risk-reducing surgeries, we encountered two samples that were apparently malignant: LGS-2 (clinically diagnosed) and OSE(P)7 (indicated by organoid characterization, **Extended Data Figure 3e,f**). Interestingly, unsupervised hierarchical clustering of gene expression data grouped these organoid lines together, thus implying biological similarity. Both organoid lines were derived from patients at high risk of developing HGS tumors. Therefore, these samples potentially represent an early time point in HGS development. Establishing and

analyzing additional early/premalignant organoid lines from pBSO material might substantiate this hypothesis and provide a unique opportunity to study early HGS tumor development.

An additional experimental platform, recently described to model colorectal cancer development[122,123,247,248], can be established through CRISPR-mediated mutation of tumor driver genes in normal organoids. Indeed, we demonstrate that normal FT organoids from OC high-risk donors can be efficiently CRISPR–Cas9 genome edited and clonally expanded afterwards, demonstrating the feasibility of such an approach in OC.

HGS tumors are frequently sensitive to platinum-based chemotherapy, whereas non-HGS tumors (such as LGS and MC tumors) are characterized by relative chemoresistance[249–252]. Consistent with these clinical observations, most HGS organoids were sensitive to platinum-based treatments, whereas non-HGS organoids (that is MBT, SBT and LGS) were more resistant (**Figure 6b**). In one case, we compared drug responses in matched organoid lines derived from primary chemosensitive (HGS-1) and recurrent chemoresistant (HGS-1-R3) tumors of a single patient. This experiment confirmed an increased resistance of the organoid line derived from the recurrent tumor to platinum-based chemotherapy, anecdotally substantiating the clinical relevance of OC organoids. Increasing the number of matched primary/recurrent organoid pairs is currently ongoing. The individual drug responses of OC organoids (for example, compare HGS-23-and HGS-3.1) illustrates the complexity of choosing the right treatment. We provide proof of concept that *in vivo* drug sensitivity of OC organoid can be tested following xenotransplantation.
In summary, we present a new organoid culture-based platform for the study of OC that supports efficient derivation and long-term *in vitro* expansion of a wide variety of OC subtypes. This living OC organoid biobank—available to the research community—faithfully recapitulates OC hallmarks, can be subjected to genetic manipulations and to drug screening and opens the door to many avenues of OC research.

METHODS

APPROVAL OF STUDIES INVOLVING HUMANS AND PATIENT-INFORMED CONSENT

The collection of patient data and tissue for the generation and distribution of normal FT, OSE and OC organoids was performed according to the guidelines of the European Network of Research Ethics Committees (EUREC) following European, national and local law. The medical ethical committee UMC Utrecht (METC UMCU) approved the

biobanking protocol: 14-472 HUB-OVI. All patients participating in this study signed informed consent forms and could withdraw their consent at any time. Available organoids are cataloged at www.hub4organoids.eu and can be requested at info@hub4organoids.eu. Distribution of organoids to third parties will have to be authorized by the METC UMCU at request of the HUB to ensure compliance with the Dutch 'medical research involving human subjects' act.

OC TISSUE PROCESSING

On arrival, OC tissues were cut into 3–5 mm^3 pieces (**Extended Data Figure 1a**). Two or three random pieces were snap frozen and stored at −80 °C for DNA isolation, two random pieces were fixed in formalin for histopathological analysis and immunohistochemistry, and the remainder were processed for organoid derivation. For organoid derivation: tissue was minced, washed with 10 ml AdDF+++ (Advanced DMEM/F12 containing 1x Glutamax, 10 mM HEPES and antibiotics). We let big tissue pieces to sink to the tube bottom with gravity (for 2–5 min), collected the supernatant and centrifuged at 1,000 r.p.m. for 5 min. In case of a visible red pellet, erythrocytes were lysed in 2 ml red blood cell lysis buffer (Roche, 11814389001) for 5 min at room temperature followed by an additional wash with 10 ml AdDF+++ and centrifugation at 1,000 r.p.m. Remaining big tissue pieces were digested in 5–10 ml AdDF+++ supplemented with 5 µM RHO/ROCK pathway inhibitor (Abmole Bioscience, Y-27632) containing 0.5–1.0 mg ml^{-1} collagenase (Sigma, C9407) on an orbital shaker at 37 °C for 0.5–1.0 h. The digested tissue suspension was sheared using 5 ml plastic pipettes. Suspension was strained over a 100 µm filter and large tissue pieces entered a subsequent digestion and shearing step. Suspension was centrifuged at 1,000 r.p.m. and the pellet was resuspended in 10 ml AdDF+++ and centrifuged again at 1,000 r.p.m. Once again, in case of a visible red pellet, erythrocytes were lysed in 2 ml red blood cell lysis buffer for 5 min at room temperature followed by an additional wash with 10 ml AdDF+++ and centrifugation at 1,000 r.p.m. Ascites/pleural effusion samples were centrifuged at 1,000 r.p.m. and treated with 2 ml red blood cell lysis buffer for 5 min at room temperature. Following erythrocyte lysis, 10 ml AdDF+++ was added and suspension was centrifuged at 1,000 r.p.m. Following removal of a large part of the ovarian stroma and the surrounding muscle layers of FT, ovary and FT samples were processed as above.

ORGANOID CULTURE

The cell pellet was suspended in 10 mg ml^{-1} cold Cultrex growth factor reduced BME type 2 (Trevigen, 3533-010-02) and 40 µl drops of BME cell suspension were allowed to solidify on pre-warmed 24-well suspension culture plates (Greiner, M9312) at 37

°C for 30 min. On BME stabilization, 500 ml of appropriate organoid medium (OC/OCwnt/OSE/FT medium, see **Supplementary Table 2**) was added and plates transferred to humidified 37 °C/5% CO2 incubators. In some cases, 25 ng ml^{-1} HGF (Peprotech) was added to the medium (**Supplementary Table 3**). Medium was changed every 3–4 d and organoids were passaged every 1–4 weeks. Organoid passaging: organoids were mechanically sheared through P1000 pipet tip connected to P200 pipet tip without a filter. Dense organoids that were not easily sheared mechanically were collected with 1 ml pre-warmed (37 °C) Accutase solution (A6964, Sigma), incubated for 1–5 min at room temperature and mechanically sheared as before. Following the addition of 10 ml AdDF+++ and centrifugation at 1,200 r.p.m, organoid fragments were resuspended in cold BME and reseeded as above at suitable ratios (1:1 to 1:4) allowing the formation of new organoids. In some lines, organoids repeatedly appeared floating in medium. These organoid lines could be transferred to repellent plates (Greiner, 662970) and expanded with medium containing 5% BME (**Supplementary Table 3**). Genetically manipulated FT clones were expanded in OCwnt medium.

SCANNING ELECTRON MICROSCOPY

To remove BME, organoids were collected with Cell Recovery Solution (Corning) and gently shacked using tube rotator, for 30 min at 4 °C. Organoids were allowed to settle down with gravity, the recovery solution was removed and 1 ml of 1% (v/v) glutaraldehyde (Sigma) in PBS was added. Following an overnight fixation at 4 °C, organoids were transferred onto 12 mm poly-l-lysine coated coverslips (Corning). The organoids were serially dehydrated by consecutive 10 min incubations in 2 ml of 10% (v/v), 25% (v/v) and 50% (v/v) ethanol-PBS, 75% (v/v) and 90% (v/v) ethanol-H2O (2x) followed by 50% ethanol hexamethyldisilazane (HMDS) and 100% HMDS (Sigma). Coverslips were removed from the 100% HMDS, air dried overnight at room temperature and mounted onto 12 mm specimen stubs (Agar Scientific). Following gold coating to 1 nm using a Q150R sputter coater (Quorum Technologies) at 20 mA, samples were examined with a Phenom PRO table-top scanning electron microscope (Phenom-World).

HISTOLOGY AND IMAGING

Tissue and organoids were fixed in 4% paraformaldehyde followed by dehydration, paraffin embedding, sectioning and standard HE staining. For the blind test, sections were randomized and analyzed by an OC pathologist. Immunohistochemistry was performed using antibodies as specified in **Supplementary Table 9**. Images were acquired on a Leica Eclipse E600 microscope and processed using the Adobe Creative Cloud software package. For time-lapse imaging, organoids were plated in BME in glass-bottom

96-well plates and mounted on an inverted confocal laser scanning microscope (Leica SP8X), which was continuously held at 37 °C and equipped with a culture chamber for overflow of 6.0% CO2. Over 16–20 h, approximately 10 H2B-mNeon-expressing organoids were imaged simultaneously in XYZT-mode using a x40 objective (NA 1.1), using minimal amounts of 506 nm laser excitation light from a tunable white light laser. Images were taken at 4 min intervals.

GENOMIC ANALYSIS

For karyotyping, 0.1 μg ml^{-1} colcemid (Gibco, 15212012) was added to the complete growth medium. About 12 h later organoids were harvested, trypsinized into single cells, incubated in hypotonic 75 mM KCl solution for 10 min and fixed in methanol:acetic acid solution (3:1). Metaphase spreads were prepared, mounted with DAPI-containing Vectashield, imaged on a DM6000 Leica microscope and quantified by manual chromosome counting. A minimum of 14 spreads was analyzed for each line. For DNA isolation, library preparation and WGS, organoid and blood samples were processed by using the DNeasy Qiagen kit. DNA from tumor tissue was isolated with the Genomic Tip Qiagen kit, supplemented with RNase treatment. Quality and quantity of samples were checked with Qubit (DNA BR). DNA integrity and RNA contamination was assessed by using Tapestation DNA screens (Genomic screen) and Nanodrop (260/280 ratio). Per sample, 500–1,000 ng of DNA was used for DNA library preparation, and whole-genome paired-end sequencing (2x150 bp) was performed on Illumina HiSeq X Ten and NovaSeq 6000 to an average coverage of 42x. **Supplementary Table 10** provides a list of all commercial and custom code used for data collection and analysis including: name, version, source and link. WGS data were processed using our in-house Illumina Analysis Pipeline (IAP) v. 2.5.1 (https://github.com/UMCUGenetics/IAP). Briefly, reads were mapped against the human reference genome GRCh37 using Burrows–Wheeler Alignment with maximal exact matches (BWA-MEM), v. 0.7.5a-r405[253]. Read mapping was followed by marking of duplicates, and indel-realignment, according to best practice guidelines[254] by the Genome Analysis ToolKit (GATK) v.3.4-46[255]. Normal cell contamination in tumor and organoid samples was estimated in silico using PURPLE v. 2.14[29]. Somatic SNVs and indels were called in the tumor and the organoids independently using the corresponding blood sample as a reference and four different tools: Strelka, v.1.0.14[256]; Varscan, v.2.4.1[257]; Freebayes, v.1.0.2[258]; and Mutect, v.1.1.7[259]. The functional effect of the somatic SNVs and indels were predicted using SnpEff v.4.1[260]. Tumor/organoid pair VCF files were then merged by selecting high-confidence SNVs and indels with a minimum alternative allele read depth of five in the tumor or ten in the organoids and called by at least two independent somatic callers in either of the samples. In addition, high-confidence

SNVs that were only detected in either the tumor or the organoid sample of a pair were called in the corresponding sample (tumor or organoid) when supported by more than 5% of the reads covering that position. CNV was detected for each sample independently using Control-FREEC, v. 7.2[261] and assuming a ploidy of 2. For sample HGS-6, a ploidy of 3 was assumed for the plots. Structural variation calling was performed using Manta, v.0.29.5[262]. For increased sensitivity, we ran Manta in the four available analysis types: singlesample, multi-sample, tumor-only and tumor-normal. When comparing SVs called in one of the tumor/organoid pairs with the matching sample, we inspected the output of the tumor-normal mode of the pertinent tumor/organoid sample with the results of the four calling modes for the matching tumor/organoid sample. Somatic variant calling could not be performed for samples without matching reference DNA (CCC-1 and END-1). In these cases, germline variant calling was performed jointly for tumor and organoid samples using GATK's Haplotype Caller, v3.4-46[255]. Germline calls were filtered against the Genome of the Netherlands (GoNL)[263] and the 1000 Genomes[264] and only variants with a predicted 'moderate' or 'high' effect (SnpEff v.4.1[260]) were kept. For SV calling of the CCC-1 and END-1 samples, the tumor-normal mode of Manta could not be used, but all other Manta variant calling workflows were performed (tumor-only, singlesample, multi-sample). To enrich for somatic SVs, only SVs larger than 10 Kb and not found in the GoNL or 1000 Genomes studies were considered for these two samples.

SINGLE-CELL WGS LIBRARY PREPARATION

Cells were sorted into 384-well plates with 5 µl of mineral oil (Sigma-Aldrich). After sorting cells, can be stored at –20 °C. Five-hundred nanoliters of lysis mix (0.001 U μl^{-1} Qiagen Protease in NEB Buffer 4) was added to each well and lysis was performed at 55 °C overnight followed by heat inactivation for 20 min at 75 °C and for 5 min at 80 °C. Five-hundred nanoliters of Restriction Enzyme mix (1 U µl–1 NLAIII in NEB Cutsmart buffer) was added to each well and restriction was performed for 3 h at 37 °C followed by heat inactivation for 20 min at 65 °C. One-hundred nanoliters of 1 µM barcoded double-stranded NLAIII adapter was added to each well. Ligation mix (1,100 µl, 182 U μl^{-1} T4 DNA Ligase in 1x T4 DNA Ligase buffer supplemented with 3 mM ATP) was added to each well and ligation was performed overnight at 16 °C. After ligation, single cells were pooled and library preparation was performed as described in Muraro *et al.*[265]. Libraries were sequenced on an Illumina Nextseq500 with 2 x 75-bp paired-end sequencing.

SINGLE-CELL WGS DATA ANALYSIS

Reads were aligned to GRCh38 using Burrows– Wheeler Aligner v0.7.14 mapping tool with settings 'bwa mem –M'[266]. Data were binned in 1 MB bins and normalized to the

expected NLAIII mappability per bin. The expected NLAIII mappability per bin was calculated by generating 108 reads from the reference genome, with every read starting at a NLAIII site. These reads were subsequently mapped and binned using the same procedure as for the experimental data. The number of reads per bin was then divided by the average number of reads per bin to acquire the expected NLAIII mappability for each bin. Regions where the expected NLAIII mappability was <0.9 or >1.2 were excluded from further analysis. After this the cells were filtered and only cells with >20,000 reads were kept for further analysis. The median read count of each cell was then set to 2 to represent a diploid genome. Data were log2 transformed to obtain log2 CN ratios and smoothed using a running mean (R package caTools) with a width of 20 MB. To remove additional low-quality cells, the variance across the genome was calculated for each cell and cells with a variance >0.3 were removed. For two dimensional visualization of the data, we first performed independent component analysis (ICA) (R package fastICA) followed by t-stochastic neighbor embedding (t-SNE) (R package Rtsne). Clustering was performed using ward.D2 hierarchical clustering on the Manhattan distances of the ICA-transformed data. Subsequently, the average copy number profile per cluster was calculated using the R package DNAcopy. Finally, a tree was constructed using ward. D2 hierarchical clustering on the manhattan distances of the DNAcopy-derived CNV profiles of the non-diploid clusters.

RNA-SEQ ANALYSIS

RNA was isolated from organoids with Trizol Reagent (Ambion). RNA libraries were generated with the Truseq Stranded Ribo-zero Sample preparation kit. RNA integrity was assessed by Tapestation (RNA screen) and quantified by Qubit (RNA). Libraries were multiplexed and paired-end sequenced (2 x 75 bp) on Illumina NextSeq. **Supplementary Table 10** provides a list of all commercial and custom code used for data collection and analysis including: name, version, source and link. RNA-seq data were processed with our in-house RNA analysis pipeline (v.2.3.0, https://github.com/UMCUGenetics/RNASeq). Reads were aligned to the human reference genome GRCh37 using STAR v. 2.4.2[267], and then read count was performed with HTSeq-count, v. 0.6.1[268]. Features (ENSEMBL definitions GRCh37, release 74) with zero read counts were filtered out (21,711 features out of 63,677). Gene symbols were mapped to the ENSEMBL features using the biomaRt package v. 2.26.1[269], and features without corresponding gene symbols and with duplicate mappings were removed. The final count matrix consisted of 30,080 rows (genes). The DESeq2 package, v1.10.1[270] was then used to normalize the read counts using the median-of-ratios method. Spearman correlation between samples was calculated using the normalized read counts from all 5,000 most variable genes and samples were clustered using hierarchical clustering with complete linkage on the correlation matrix. The genetically modified organoid lines were analyzed using the same DESeq2 pipeline.

METHYLATION ANALYSIS

For methylation analysis 210 ng of genomic DNA was used. DNA was sodium bisulfite converted with the Zymo Research EZ DNA methylation kit (Zymo Research) and treated with the InfiniumHD FFPE Restore kit (Illumina). Next, the DNA was hybridized to the Infinium MethylationEPIC 850 K BeadChip (Illumina) to analyze the genome-wide methylation status of 865,859 methylation sites. **Supplementary Table 10** provides a list of all commercial and custom code used for data collection and analysis including: name, version, source and link. For methylation data analysis, fluorescence intensity data (.IDAT) files were analyzed by using the minfi R package[271]. Beta-values were extracted after applying a normalization step with minfi preprocessFunnorm. Pearson correlation of beta-values between samples was calculated, and subsequently unsupervised hierarchical clustering of correlation values was performed on the 11,720 most variable probes.

GENE EDITING

Organoids derived from early passaged (P0–P3) FT organoids were dissociated into small clumps using pre-warmed Accutase solution (A6964, SIGMA), washed once with AdDF+++ and twice with Opti-MEM (11058021, Life technologies). Cells were suspended with 100 µl Opti-MEM containing RHO/ ROCK pathway inhibitor (10 µM) and 10 µg of pSpCas9(BB)-2A-GFP (a gift from F. Zhang[272] from the Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA, USA), Addgene plasmid no. 48138) with guide RNA (gRNA) targeting TP53 (GACGGAAACCGTAGCTGCC)[123] or combination of gRNA targeting TP53 and RB1 (GTTCGAGGTGAACCATTAAT) genes, and transferred into 2 mm gap NEPA electroporation cuvette (lot no. 2S1509). For electroporation, we utilized NEPA21 type-II electroporator (**Supplementary Table 11**). Following electroporation, 300 µl of complete growth medium was added to the cells and they were incubated at room temperature for 15 min. Cells were centrifuged, suspended in 200 µl BME and plated as previously described. Complete medium was added after cell BME suspension drops had solidified. Two to three days after electroporation, 10 µM nutlin-3 (Cayman Chemical) was added to the growth medium. Two to three weeks after electroporation, single organoids were picked and transferred into 1.5 ml microcentrifuge tubes containing 200 µl of pre-warmed Accutase. Following 2–3 min incubation, organoids were sheared into small cell clumps by pipetting, washed with 1 ml AdDF+++ and centrifuged for 5 min at 2,000 r.p.m. Cells were resuspended with 40 µl BME and plated. For genotyping, genomic DNA was isolated using Viagen Direct PCR (Viagen). GoTaq Flexi DNA polymerase (Promega) was used for PCR amplification. Primer sequences: P53_for, 5'-CAGGAAGCCAAAGGGTGAAGA-3'; P53_rev,

5'-CCCATCTACAGTCCCCCTTG-3'; RB1_for, 5'-CAGAGTAGAAGAGGG ATGGCA-3'; RB1_rev, 5'-CAGTGATTCCAGAGTGACGGA-3'. Products were cloned into pGEM-T Easy vector system I (Promega) and sequenced using T7 sequencing primer.

LENTIVIRUS TRANSDUCTION OF ORGANOIDS

To visualize mitoses, organoids were infected with lentivirus encoding mNeon-tagged histone-2B and a puromycin resistance cassette (pLV-H2B-mNeon-ires-Puro[123]) as previously described[273].

DRUG SCREEN AND VIABILITY ASSAY

Dispase II (1 mg ml–1; Invitrogen) was added to the medium of the organoids and these were incubated for 10 min at 37 °C to digest the BME. Subsequently, organoids were mechanically dissociated by pipetting and were filtrated using a 70 mm nylon cell strainer (Falcon), resuspended in 2% BME/ growth medium (15,000–20,000 organoids ml–1) before plating in 50 µl volume (Multi-drop Combi Reagent Dispenser) on BME pre-coated 384-well plates. The drugs and their combinations were added 1 h after plating the organoids using the Tecan D300e Digital Dispenser. Drugs were dispensed in a randomized manner and DMSO end concentration was 1% in all wells. 120 h after adding the drugs, ATP levels were measured using the Cell-Titer Glo2.0 (Promega BV) according to the manufacturer's instructions and luminescence was measured using a SpectraMax microplate reader (Molecular Devices). Results were normalized to vehicle (DMSO=100%) and baseline control (navitoclax 20 µM). Data were analyzed using GraphPad Prism 6. Using the trapezoid rule for numerical integration, the AUC was approximated between the lowest and highest concentrations screened in the actual assay. Organoid drug sensitivity was represented by the average AUC of two technical replicates and independent experimental repetitions in a subset of treatments and visualized using RStudio. Experimental repetition with a subset of drugs was performed in the following lines: FT-1, FT(P)-1, END-1.1, END-1.2, MC-2.1, HGS-1, HGS-1-R2, HGS-3.1, HGS-3.2, HGS-22, HGS-23. Euclidean distance between samples was measured using the normalized (row Z-score) AUC. Alpelisib (BYL719), catalog no. S2814, Selleckchem; adavosertib (MK-1775), catalog no. S1525, Selleckchem; AZD8055, catalog no. S1555, Selleckchem; carboplatin, catalog no. S1215, Selleckchem; gemcitabine, catalog no. S1714, Selleckchem; MK-2206, catalog no. S1078, Selleckchem; niraparib (MK-4827), catalog no. S2741, Selleckchem; nutlin-3, catalog no.10004372, Cayman Chemical; paclitaxel, catalog no. S1150, Selleckchem; pictilisib (GDC-0941), catalog no. S1065, Selleckchem.

RECAP ASSAY

Organoids were incubated at 37 °C/5% CO2 humidified atmosphere and an equal number of organoids were transferred to 3 cm Petri dishes containing 2 ml of medium. One petri dish was irradiated with 5 Gy X-rays (200 kV, 4 mA, YXLON Y.TU 225-D02) and the other Petri dish was mock-treated (that is not irradiated). EdU (0.02 mM; ThermoFisher Scientific, Click-iT EdU Alexa Fluor 647 Imaging Kit, catalog no. C10340) was added to the organoids and incubated for 2 h at 37 °C/5% CO2 humidified atmosphere on a 60 r.p.m. rotating platform. The organoids were transferred to 15 ml falcon tubes and after the organoids were settled down by normal gravity at room temperature, medium was removed and replaced by 10 ml buffered formalin (10%). Organoids were fixed for 1 h on a rotating device at room temperature, washed twice with PBS and stored in 70% ethanol at 4 °C. The organoids were embedded into paraffin, sliced into 5 µm slices and incubated in 60 °C o/n on StarFrost microscope slides (76 x 26 mm, Knittel glass). Immunofluorescence staining was performed to stain for DAPI (ThermoFisher Scientific, catalog no. P36935,), geminin (primary antibody rabbit, Proteintech Europe, catalog no. 10802-1-AP), RAD51 (primary antibody mouse, Gene Tex, GTX70230) and EdU (ThermoFisher Scientific, Click-iT EdU Alexa Fluor 647 Imaging Kit, catalog no. C10340). RAD51 foci were scored blindly in 10 randomly chosen organoids, counting at least 100 geminin+ cells in total for both the irradiated and the non-irradiated organoids. Biological repetitions were done as indicated in figure legend (**Figure 6**). A nucleus was scored as RAD51 positive if it contained more than five foci. Organoids in which less than six cells were counted as geminin+ were filtered out from the analysis.

ORGANOID-DERIVED XENOGRAFT

Experiments on NSG mice were carried out at the Netherlands Cancer Institute according to local and international regulations and ethical guidelines, and were approved by the local and central animal experimental committee at the Netherlands Cancer Institute (AVD3010020172464; IVD 9.1 EGP 8102) 8102). Ovarian injection: mice were anesthetized with isoflurane (3% induction, and 2% maintenance) and a small incision in the flank and peritoneum was made. The ovarium was gently taken from the abdominal cavity and tumor cells are slowly injected with an insulin needle (Terumo 29 G x 1/2, 0.33 x 12 mm) into the bursa. The ovarium was positioned back in the abdominal cavity, and peritoneum and skin were sutured separately. IVIS-imaging: mice were injected with 10 µ1 per g body weight of Beetle luciferin (promega E1605) and after 10 min bioluminescence was measured on the IVIS Lumina. After the mice were killed, the ovarium was taken out and embedded in paraffin for further analysis.

Intervention study: experiments on NSG mice were carried out at the Netherlands Cancer Institute according to local and international regulations and ethical guidelines, and were approved by the local animal experimental committee at the Netherlands Cancer Institute (AVD301002015407; IVD 1.1 EGP 8583). Subcutaneous injection: mice are subcutaneously injected with the organoid lines. Caliper measurements were performed three times per week. When the tumors reached a size of 50 mm3, treatment started with either Vehicle (saline) or Gemcitabine (2 mg kg), intraperitoneal injection 5 times per week (5 on, 2 off) for 4 consecutive weeks. Ten mice per treatment arm were included. Tumor size was monitored for 55 d; mice that died before that time point (after surgery or gemcitabine treatment) were excluded from the analysis.

STATISTICAL ANALYSES

Where applicable, statistical methods are outlined in the respective figure legends. Statistical analysis was performed utilizing Microsoft Excel, GraphPad and R package. P values were calculated using a two-tailed Student's t-test. DNA and RNA sequencing analysis details can be found in the relevant Methods sections. For karyotyping a minimum of 14 metaphase spreads was analyzed for each line. For single-cell DNA analysis 791 cells from 2 recurrent tumor samples and 3 corresponding organoid lines were analyzed. Drug screen killing curves show the average ± s.e.m. of two technical replicates. AUC of independent drug screen repetitions was averaged and presented in a drug sensitivity heat map (experimental repetitions (n=2) at different passage numbers in a subset of treatments was carried out in 11 independent organoid lines, **Extended Data Figure 5d**). For animal intervention experiments, 10 mice per treatment arm were included. Mice that died before the experimental end-point were excluded from analysis. In the case of representative results, the number of independent organoid lines or experimental repetitions and their relevant description are indicated in the figure legend.

CLINICAL DATA

Patients agreed with the use of their clinical data by signing informed consent. Clinical data was extracted from the patient file by the Dutch Cancer Registration and included age at diagnosis, patient history, BRCA mutation status, tumor characteristics and treatment modalities.

REPORTING SUMMARY

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

BAM files for DNA and RNA sequencing data are made available through controlled access at the European Genome-phenome Archive (EGA) which is hosted at the EBI and the CRG (https://ega-archive.org), under accession number EGA: EGAS00001003073. Data access requests will be evaluated by the UMCU Department of Genetics Data Access Board (EGAC00001000432) and transferred on completion of a material transfer agreement and authorization by the medical ethical committee UMCU at request of the HUB to ensure compliance with the Dutch 'medical research involving human subjects' act.

CODE AVAILABILITY

Illumina data processing pipeline v2.2.1 is available at https://github.com/UMCUGenetics/IAP/releases/tag/v2.2.1 and RNA analysis pipeline v2.3.0 is available at https://github.com/UMCUGenetics/RNASeq. All other custom code used for this study is available at https://github.com/UMCUGenetics/ OvCaBiobank

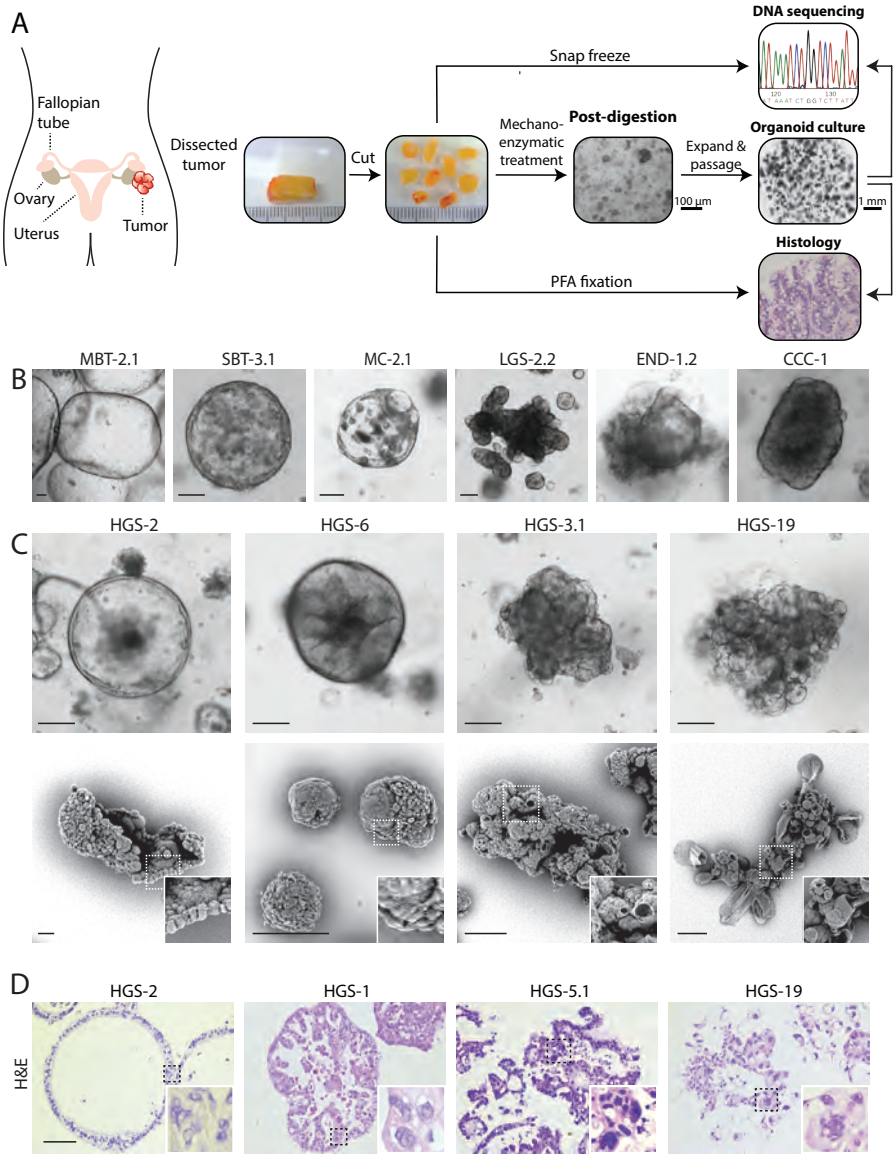
LIST OF SUPPLEMENTARY DATA

ED=extended data

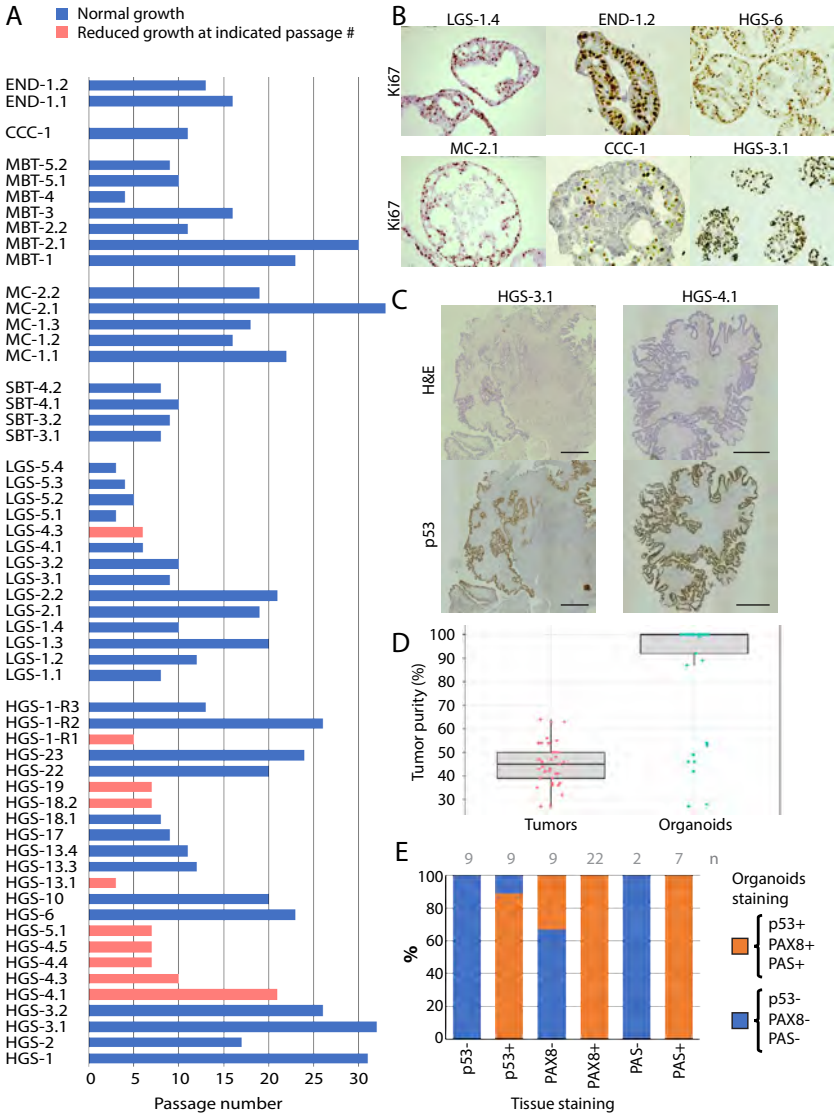
ED Figure 1.	Derivation and morphological differences of OC organoids.
ED Figure 2.	Organoid passage number overview and normal cell contamination in tumors and organoids.
ED Figure 3.	FT and OSE organoids.
ED Figure 4.	Genome-wide tumor and organoid pair comparison.
ED Figure 5.	Molecular characterization, drug screening and xenografts of OC organoids.
ED Figure 6.	CRISPR–Cas9 mediated genetic manipulation in FT organoids.

*Table S1.	Patient clinical data.
*Table S2.	Medium recipe.
*Table S3.	OC organoid line information.
*Table S4.	Organoid subtype diversity and derivation efficiency
*Table S5.	FT and OSE organoid information
*Table S6.	OC organoid histological analysis
*Table S7.	RNA and DNA sequencing related information of OC organoids
*Table S8.	Organoid derived xenografts
*Table S9.	Antibody list
*Table S10.	Code Availability
*Table S11.	Electroporation setup

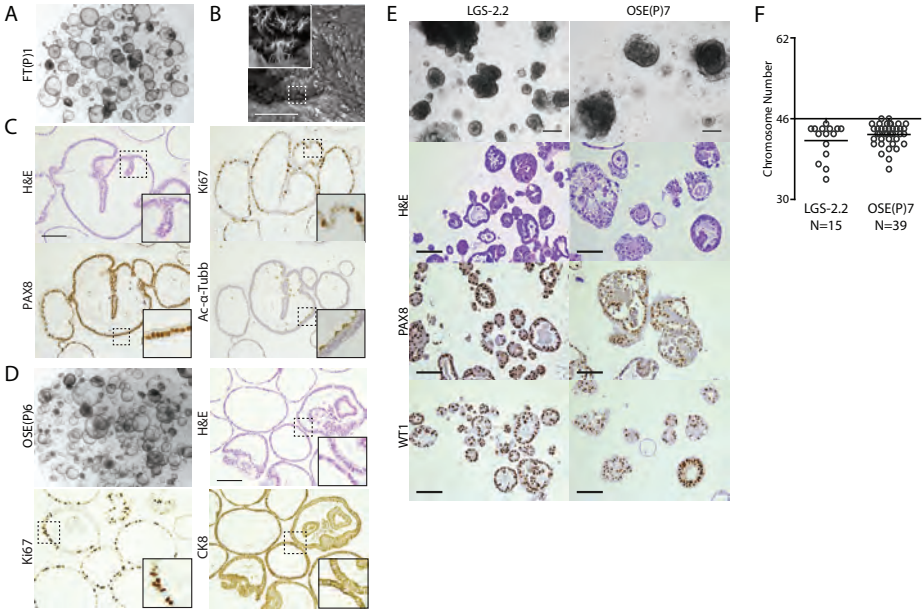
**Table S1-11 are available online at: https://tinyurl.com/Ch2Suppl or scanning the QR code below*



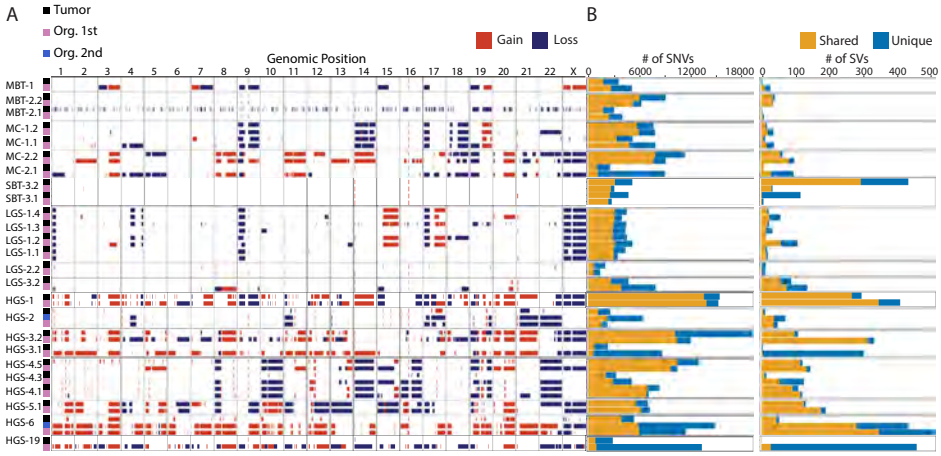
Extended Data Figure 1: Derivation and morphological differences of OC organoids. (**A**) Schematic of OC organoid derivation. (**B**) Bright-field images of MBT, SBT, MC, LGS, END and CCC organoids (left to right), depicting different organoid morphologies. Morphological description of 50 independent organoid lines is provided in Supplementary Table 6. Scale bar, 100 µm. (**C**) Bright-field (top) and SEM (bottom) images demonstrating main morphologies among different HGS organoid lines. Starting with cystic and well-organized cellular polarity, where microvilli are directed toward the organoid lumen (most left) to dense organoids that gradually (from left to right) show reduced circularity and cellular cohesiveness up to a grape-like shape morphology (most right). Scale bar, 100 µm. **v** High-magnification H&E staining images displaying representative examples of HGS organoid morphologies as well as nuclear and cellular atypia, typically displayed by HGS tumors. Histological description of 50 independent organoid lines is provided in Supplementary Table 6. Scale bar, 100 µm.



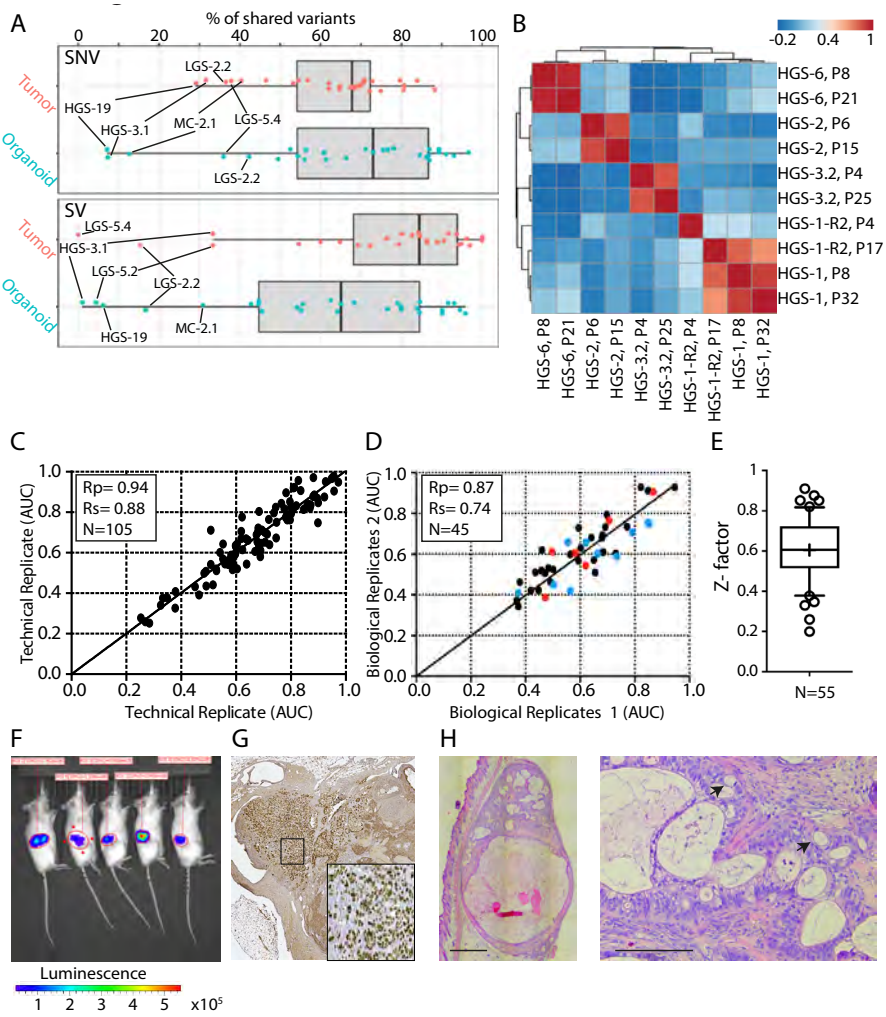
Extended Data Figure 2: Organoid passage number overview and normal cell contamination in tumors and organoids. (**A**) Column bar graph depicting organoid maximum passage number up until the moment of submission. Organoids that stopped/slowed down their growth are indicated in orange. (**B**) Representative images of Ki67 staining of six independent organoid lines show a high percentage of ki67-positive proliferating cells. (**C**) Histological and immunohistochemical images of tumor tissue (derived from two independent patients) showing tumor cell purity within different samples, based on H&E and p53 staining. Scale bar, 0.5 mm. (**D**) Tukey box-and-whisker plot (1.5x interquartile range) presenting bioinformatic estimation of tumor cell purity percentage of both tissue (n=35) and organoid (n=36) based on WGS data using PURPLE. Horizontal bars represent the median of all dots. Mean and standard deviation across all samples are as follows: 45 ± 9.2% (tissue) and 88.1 ± 23% (organoids). (**E**) Stacked bar chart showing the percentage of organoid lines that are positive for p53, PAX8 and periodic acid–Schiff (PAS) staining (orange) and negative (blue) grouped per original tumor staining status (see also Supplementary Table 6). Total number (n) of tissues stained per group are indicated.



Extended Data Figure 3: FT and OSE organoids. **(A)** An overview image of normal FT organoids embedded in 40 µl BME drops, displaying a cystic morphology. All FT organoid lines that were established (n=22) displayed cystic morphology. **(B)** Representative SEM image showing ciliated cells facing FT organoid lumen. Scale bar, 50 µm. SEM was performed on one FT organoid line. **(C)** Histological analysis of FT organoids demonstrating H&E, Ki67, PAX8 and Ac-α-tubb staining. Histological analysis was performed on three independent FT organoid lines with similar results. Scale bar, 100 µm. **(D)** An overview image of normal OSE organoids embedded in 40 µl BME drops displaying cystic morphology (top left image). Seven out of eight OSE organoid lines that were established displayed cystic morphology. OSE organoid images of H&E, Ki67 and cytokeratin 8 (CK8) staining, demonstrating a cystic morphology of proliferative epithelial cells. Histological analysis was performed on two independent OSE organoid lines with similar results. Scale bar, 100 µm. **(E)** First row: bright-field images of LGS-2.2 (left) and OSE(P)7 (right) organoid lines. Unlike normal FT and OSE that display cystic morphology both lines show dense phenotype. OSE(P)7 is the only OSE organoid line that displays dense phenotype. Scale bar, 200 µm. Second to last rows: histological and immunohistochemical images demonstrate that organoids are positively stained for PAX8 and WT1, markers of OC serous subtypes. Organoids display reduced cellular organization in comparison to normal FT and OSE organoids. Scale bar, 100 µm. **(F)** Scatter plot presenting metaphase spread analysis and mean for each line. Both lines present aneuploidy.

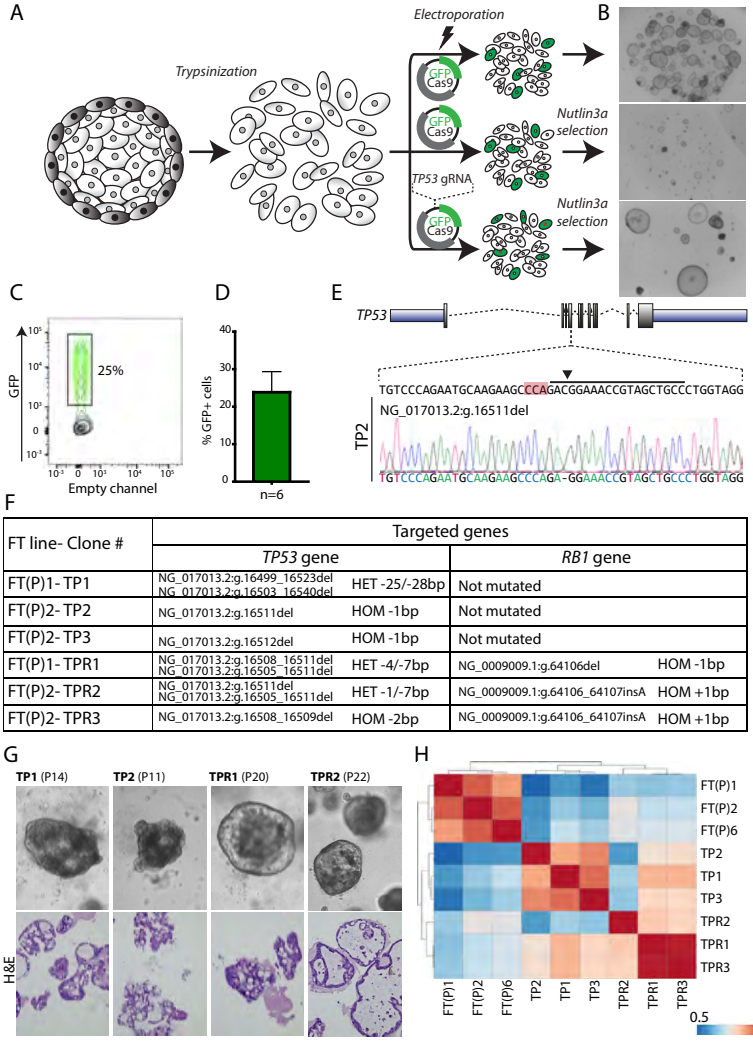


Extended Data Figure 4: Genome-wide tumor and organoid pair comparison. **(A)** Genome-wide CNVs in tumor/organoid pairs (black, tumors; pink, organoids early passage; blue, organoids late passage) depicting gains (red) and losses (blue). **(B)** Number of shared (yellow) and unique (blue) SNVs (on the left) and SVs (on the right) between tumor/organoid pairs. Shared variants are those that can be found in the corresponding paired sample. Passage number at which organoid lines were sequenced is given in Supplementary Table 7.



Extended Data Figure 5: Molecular characterization, drug screening and xenografts of OC organoids. **(A)** Tukey box and whisker plot (1.5x interquartile range) summarizing the percentage of shared variants across all tumor (red) and organoid (green) samples. Right and left panels display SNVs and SVs, respectively. Horizontal bars represent the median of all dots. Mean and standard deviation across all samples are as follows: SNVs, 82.95 ± 8.18% (tissue, n=31) and 75.62 ± 23.13% (organoids, n=31); SVs, 78.14 ± 22.11% (tissue, n=31) and 60.47 ± 29.13% (organoids, n=31). Samples with a low percentage of shared variants are indicated. **(B)** Heat map of five independent organoid lines from both early and late passages based on 11,720 methylation probes. The heat map colors represent Pearson correlation values, as calculated from the methylation beta-values. Clustering of the correlation values was performed using hierarchical clustering based on complete linkage. **(C)** Scatter plot of AUC values across all drug screening data, displaying high correlation between technical replicates (Pearson correlation=0.94, R2=0.88, n=105). **(D)** Scatter plot of AUC values of biological replicates, displaying high correlation (Pearson correlation=0.87, R2=0.74, n=45). Colored dots represent biological replicates in which passage differences between experimental repetition is as follows: 1–2 passages, n=29 (black); 3–5 passages, n=10 (blue) and 13–22 passages, n=6, (red), demonstrating stable drug sensitivity even after prolonged passaging. **(E)** Box-and-whisker plot (10th–90th percentile) showing Z-factor distribution and mean across all drug screening plates. Mean=0.61, ranging between 0.2 and 0.91, n=55. **(F)** Bioluminescence imaging of mice, orthotopically transplanted with

luciferase expressing organoid lines depicting tumor growth. A summary of organoid-derived xenograft experiments is presented in Supplementary Table 8. **(G)** p53 staining of organoid derived xenograft (HGS-3.1) on orthotopic transplantation into the mouse bursa shows p53 overexpression in tumor cells. **(H)** Histological analysis of an organoid-derived xenograft (MC-2.1) on subcutaneous transplantation. H&E staining shows haphazardly arranged neoplastic glands lined by columnar cells with variable numbers of goblet cells (arrows), which are specific features of MC. A summary of organoid-derived xenograft experiments is presented in Supplementary Table 8. Left image scale bar, 1 mm. Right image scale bar, 200 µm.



FT line- Clone #	Targeted genes				
	TP53 gene		*RB1* gene		
FT(P)1- TP1	NG_017013.2:g.16499_16523del NG_017013.2:g.16503_16540del	HET -25/-28bp	Not mutated		
FT(P)2- TP2	NG_017013.2:g.16511del	HOM -1bp	Not mutated		
FT(P)2- TP3	NG_017013.2:g.16512del	HOM -1bp	Not mutated		
FT(P)1- TPR1	NG_017013.2:g.16508_16511del NG_017013.2:g.16505_16511del	HET -4/-7bp	NG_0009009.1:g.64106del	HOM -1bp	
FT(P)2- TPR2	NG_017013.2:g.16511del NG_017013.2:g.16505_16511del	HET -1/-7bp	NG_0009009.1:g.64106_64107insA	HOM +1bp	
FT(P)2- TPR3	NG_017013.2:g.16508_16509del	HOM -2bp	NG_0009009.1:g.64106_64107insA	HOM +1bp	

Extended Data Figure 6: CRISPR–Cas9 mediated genetic manipulation in FT organoids. **(A)** Schematic of normal FT organoid electroporation. FT organoids were dissociated into small cell clumps and electroporated with either an empty vector or a vector containing a gRNA directed against TP53. Cells were plated and after 2 d of recovery nutlin3a was added. **(B)** Overview images of organoids 2 weeks after electroporation. Organoids that were electroporated with an empty vector and not treated with nutlin3a

showed nice recovery following electroporation (top), whereas the growth of organoids electroporated in a similar manner was dramatically inhibited when nutlin3a was added (middle). Surviving clones that are not inhibited by nutlin3a treatment are visible only when organoids were electroporated with a vector containing TP53 gRNA (bottom). Four independent electroporation experiments followed by nutlin3A treatment were conducted giving rise to multiple nutlin3A resistant clones. **(C)** A representative flow cytometry analysis of organoids 48 h following electroporation demonstrating 25% of the cell express GFP. Summary of six independent repetitions of this experiment are presented in d. **(D)** Box-and-whisker plot (minimum to maximum) showing the percentage of GFP positive cells following electroporation. Horizontal bars and dashed horizontal bars represent median and mean of all dots, respectively. Mean ± s.d.=23.8 ± 5.5%, median=25.5%. Six independent experiments that were conducted with three different FT organoid lines are presented, demonstrating high and robust electroporation efficiency. **(E)** An example of CRISPR–Cas9 mediated editing of TP53 gene in FT organoids. Targeted locus is presented and gRNA (solid line), PAM sequence (red highlight) and cut point (arrow head) are indicated. Sequencing results revealed out-of-frame deletions induced by CRISPR–Cas9 editing. **(F)** Table presenting six FT genetically engineered clones derived from two independent donors (FT(P)1 and FT(P)2). For each clone, targeted gene description (in both TP53 and RB1 genes) including HGVS nomenclature is presented. (HET, heterozygous; HOM, homozygous). **(G)** BF images (top) and H&E staining (bottom) of four independent clones show deviation from cystic and well-organized normal FT organoid morphology. Passage number is indicated. This analysis was conducted on three independent TP clones (loss-of-function mutations in the TP53 gene) and three independent TPR clones (loss-of-function mutations in the TP53 and RB1 genes) with similar results. **(H)** Heat map of Spearman correlation values of three independent normal FT organoid lines (derived from different donors) and genetically engineered clones (n=3 independent TP clones (loss-of-function mutations in the TP53 gene) and 3 independent TPR clones (loss-of-function mutations in the TP53 and RB1 genes)), using RNA-seq expression data. Read counts were normalized for sequencing depth and the 1,000 most-variable genes were used. Clones were assigned into different groups according to their mutational profile.

ACKNOWLEDGEMENTS

We thank T. Bayram for support of ethical regulatory affairs. We acknowledge A. Brousali, P. van der Groep, A. Constantinides, A. Snelting and O. Kranenburg of the Utrecht Platform for Organoid Technology (U-PORT; UMC Utrecht) for patient inclusion and tissue acquisition. We thank the Integraal Kankercentrum Nederland (IKNL) and M. van der Aa for supplying clinical data, and I. Renkens for help with DNA isolations. We acknowledge E. Stelloo for her help with culturing organoids. We thank the people from the Preclinical Intervention Unit of the Mouse Clinic for Cancer and Ageing (MCCA) at the NKI for performing the intervention studies. We thank B. Artegiani and T. Dayton for critically reading the manuscript. O.K. was supported by Marie Skłodowska-Curie IF grant 658933 – HGSOC. This work was funded by the gravitation program CancerGenomiCs.nl from the Netherlands Organisation for Scientific Research (NWO), MKMD grant (114021012) from Netherlands Organization for Scientific Research (NWO-ZonMw), Stand Up to Cancer International Translational Cancer Research Grant, a program of the Entertainment Industry Foundation administered by the AACR, Dutch Cancer Society (KWF) grant UU2015-7743, Dutch Cancer Society (Alpe d'HuZes) grant EMCR 2014-7048, and a grant from the Gieskes Strijbis Foundation (1816199). The Oncode Institute is supported by the Dutch Cancer Society.

AUTHOR CONTRIBUTIONS

Conceptualization: OK, HC
Methodology: OK, HC
Software: JEVI, MR, LK, WK
Formal analysis: OK, KL, NH, JEVI, MR, TJ, PD, SAR, LK, WK
Investigation: OK, KL, CW, NH, AVB, HB, JK, SAR, LK, NP, RT, LW, BP
Resources: CW, LW, HV, MW, VH, BN, PW, MvdV, TB, KG, RZ
Data curation: OK, CW, JEVI
Writing—original draft: OK, CW, JEVI, WK, HC
Visualization: OK, CW, KL, JEVI, WK, LK, NH
Supervision: MvdV, JB, RZ, HS, WK, AO, HC
Project administration: OK, CW, WK, HC
Funding acquisition: JB, RZ, PW, WK, HC

COMPETING INTERESTS

The authors declare no competing interests.

3

PATIENT-DERIVED OVARIAN CANCER ORGANOIDS MIMIC CLINICAL RESPONSE AND EXHIBIT HETEROGENEOUS INTER- AND INTRAPATIENT DRUG RESPONSES

Patient-derived ovarian cancer organoids mimic clinical response and exhibit heterogeneous inter- and intrapatient drug responses

Chris J. de Witte[1,2], Jose Espejo Valle-Inclan[1,2], Nizar Hami[2,3], Kadi Lõhmussaar[2,4], Oded Kopper[2,4], Celien P.H. Vreuls[5], Geertruida N. Jonges[5], Paul van Diest[5], Luan Nguyen[1,2], Hans Clevers[2,4], Wigard P. Kloosterman[1], Edwin Cuppen[1,2,6], Hugo J.G. Snippert[2,3], Ronald P. Zweemer[7] , Petronella O. Witteveen[8,9,$] & Ellen Stelloo[1,2,9,10,$]

[*]These authors contributed equally to this work
[$]corresponding authors

[1]Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, the Netherlands
[2]Oncode Institute, the Netherlands
[3]Molecular Cancer Research, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, the Netherlands
[4]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences and University Medical Center Utrecht, the Netherlands
[5]Department of Pathology, University Medical Center Utrecht, Utrecht University, the Netherlands
[6]Hartwig Medical Foundation, Amsterdam, the Netherlands
[7]Department of Gynecological Oncology, Division Imaging and Oncology, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands
[8]Department of Medical Oncology, Cancer Center, University Medical Center Utrecht, Utrecht University, the Netherlands.

Adapted from: Cell Reports 31:107762 (2020);
https://doi.org/10.1016/j.celrep.2020.107762

ABSTRACT

There remains an unmet need for preclinical models to enable personalized therapy for ovarian cancer (OC) patients. Here, we evaluate the capacity of patient-derived organoids (PDOs) to predict clinical drug response and functional consequences of tumor heterogeneity. We included 36 whole-genome characterized PDOs from 23 OC patients with known clinical histories. OC PDOs maintain genomic features of the original tumor lesion and recapitulate patient response to neoadjuvant carboplatin/paclitaxel combination treatment. PDOs display inter- and intrapatient drug response heterogeneity to chemotherapy and targeted drugs, which can partially be explained by genetic aberrations. PDO drug screening identifies high responsiveness to at least one drug for 88% of patients. PDOs are valuable preclinical models that can provide insights in drug response for individual patients with OC, complementary to genetic testing. Generating PDOs of multiple tumor locations can improve clinical decision making and increase our knowledge on genetic and drug response heterogeneity.

INTRODUCTION

Epithelial ovarian cancer (OC) is characterized by the development of chemotherapy resistance and poor survival. Overall survival for patients with OC has only slightly improved over the past decades, despite developments in the field such as optimized surgical tumor resection, administration of (hyperthermic) intraperitoneal chemotherapy and introduction of targeted treatments such as PARP-inhibitors[274]. While most patients with OC respond well to initial treatment, the majority will develop recurrent disease within the first two years and become resistant to chemotherapy. In the setting of recurrent disease, a wide range of chemotherapeutic and targeted drugs is available. PARP-inhibitors are indicated for patients who experienced complete or partial response to previous platinum treatment, irrespective of *BRCA1/2*-mutation status[275]. Still, *BRCA1/2*-mutation carriers experience more benefit from PARP-inhibition compared to patients with homologous recombination (HR)-proficient tumors[275]. However, for the majority of relapsed patients and drugs, no genetic markers are available to predict response. These patients might benefit from patient-derived model systems that can be employed to test response to drugs prior to treatment in the clinic.

Traditionally, OC drug response has been studied in 2D-cell lines and xenografts. 2D-cell lines offer a relatively cheap and quick model system, suitable for high-throughput drug screening; while patient-derived xenografts offer the potential to study tumor drug response in a living organism, but are not suitable for high-throughput drug screening experiments[276]. In the past decade patient-derived organoids (PDOs) have been established[277], a 3D-cell culture model system that maintains cellular heterogeneity of healthy tissues and tumors. Recently, PDOs of OC were established which represent the genomic features of the original tumors[278–280]. Furthermore, a drug screening comparison between 2D-cultures and PDOs of OC revealed that cytostatic drug efficacy was dependent on the employed culture system; PDO drug responses correlated better with genomic aberrations compared to 2D-cell cultures[280]. To employ the organoid system to guide treatment choice in the clinic, it is vital that the correlation between PDO drug response and clinical drug response is established. To this extent, prospective clinical trials have been performed with PDOs of patients with colorectal cancer, in which *in vitro* drug screening recapitulated patient response to chemotherapy and targeted drugs[138,281]. For OC, we and others previously provided anecdotal evidence on the correlation between clinical and PDO drug response[278,279,282], but direct comparisons are still limited.

When predicting treatment response, genetic heterogeneity should be considered. Epithelial OC, especially the high-grade serous subtype, is a heterogeneous disease with widespread inter- and intrapatient genetic heterogeneity[78,283]. A virtue of the PDO model

system is the possibility to study genetic and phenotypic tumor heterogeneity[125].
In this study, we systematically assessed if *in vitro* drug response of OC PDOs correlates to patients' clinical response to chemotherapeutics. We studied inter- and intrapatient drug response heterogeneity to a wide range of chemotherapeutics as well as targeted drugs, and linked differential drug response to genetic variation.

RESULTS

PDOS CAN BE (RAPIDLY) ESTABLISHED FROM DIFFERENT OC SUBTYPES

In total, we included 36 PDOs (of which 29 were established previously[278]), derived from 23 patients with different histological subtypes of OC who underwent primary or interval debulking surgery or ascites drainage (**Table S1**). PDO sample names are informative of histological subtype as well as patient (first number) and tumor location (second number). We have previously demonstrated that PDOs are largely similar to the carcinoma fields within their matching tumor, based on histopathological assessment[278]. The majority of PDOs in our biobank were thoroughly characterized by whole-genome sequencing and histopathological examination and biobanked prior to drug testing, to establish a reliable platform. This resulted in a considerable length of time from PDO establishment to drug screening. However, in order to incorporate PDO-based drug response prediction in clinical care, PDO establishment and screening must be executed within a short time span. As a pilot experiment, we successfully established and rapidly screened organoids from a patient with recurrent disease (HG-26; **Figure S1A-C**). Within 20 days of tumor collection the response to six therapies became available.

PDOS RETAINED GENOMIC FEATURES OF THE ORIGINAL TUMOR LESION

We characterized 36 organoids, 30 matching tumors and 31 germline samples by whole-genome sequencing to an average coverage of 32X. Passage numbers at which PDOs were sequenced are provided in **Table S1**. First, we compared the genomic profiles of PDOs and the tumors they were derived from. An average of 8,290 and 10,358 SNVs were identified in the parental tumor specimen and their matched PDO, respectively. On average 67% of variants were shared between the tumor and PDO, 6% of the SNVs were unique to the tumor, and 27% to the PDO (**Figure S2A**). Assessment of CNAs demonstrated comparable copy-number states in the majority of pairs (**Figure S2B-C**). HGS-3.1, LGS-2.2 and MC-2.1 presented with a much higher number of SNVs than their matched tumor specimen and considerable CNA dissimilarities within PDO-tumor pairs. These exceptions are likely due to a high degree of normal cell contamination in

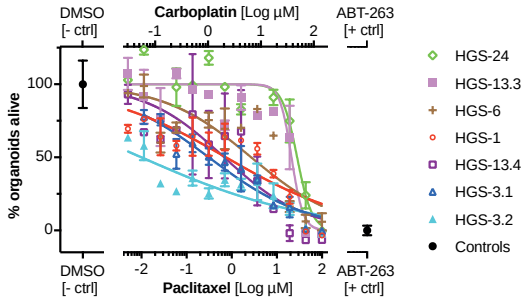
the tumor samples, which was confirmed by PURPLE, a purity ploidy estimator (**Table S7**)[284]. In general, based on PURPLE tumor content estimates (**Table S7**), PDOs are enriched for tumor cells, whereas tumor samples are heterogeneous, representing a mix of tumor cells and normal cells. For tumor samples with tumor content, PDOs retained the genomic features of the original tumor lesions.

PDO DRUG RESPONSE CORRELATES TO PATIENTS' CLINICAL RESPONSE

Next, we evaluated the potential of PDOs to reflect patients' drug response to chemotherapy. For this we selected all PDOs that were derived at interval debulking surgery from patients with high grade serous (HGS) OC, with known clinical histories (**Table S1** (clinical comparison) and **S2**). Seven PDOs (derived from five patients) were exposed to carboplatin and paclitaxel combination treatment *in vitro* and we could directly compare their response to the patient's clinical response. Related samples HGS-3.1 and HGS-3.2 were most responsive to carboplatin and paclitaxel combination treatment (AUC=0.37 and 0.29), while HGS-24 was the least responsive (AUC=0.88) (**Figure 1**). These PDO drug responses showed a statistically significant correlation ($p<0.01$) with clinical response as measured by histopathological (chemotherapy response score (CRS)), biochemical (normalization of serum biomarker CA-125) and radiological (RECIST) responses (**Figure 1; Figure S1D-H**). The derivation of organoids upon interval debulking was restricted to CRS1 and CRS2 scored samples and therefore a high-risk subgroup of patients, as CRS3 scored samples will not have macroscopic tumor lesions from which the pathologist can provide tissue for organoid derivation. PDOs derived from tumor locations with no or minimal histopathological response (CRS=1) were less responsive to carboplatin and paclitaxel combination treatment compared to PDOs derived from tumor locations with appreciable pathological response (CRS=2) (p=5.821e-05, Wilcoxon signed-rank test)[285]. Biochemically, clinical drug response is measured according to the response criteria and timing of normalization of CA-125[286,287]. Even though all patients exhibited CA-125 response during primary treatment, PDOs derived from patients who did not reach CA-125 normalization during primary treatment were less responsive to the chemotherapeutics compared to PDOs from the patient in whom CA-125 levels normalized (<35kU/L, p=0.0004). Radiological response was assessed according to the RECIST criteria (version 1.1)[287], comparing imaging data at initiation of chemotherapy to imaging data prior to interval debulking based on CT-scanning. PDOs derived from patients with RECIST stable disease were less responsive to carboplatin and paclitaxel combination treatment compared to PDOs from patients with RECIST partial response (p=0.0092). To compare long-term clinical response to PDO response, recurrence and survival were assessed. All patients experienced recurrent disease within four to 14

months after the last primary treatment, and 6-month progression-free survival (PFS) did not correlate to PDO drug response. After 17 months 50% of patients with FIGO stage IV HGS OC are still alive (**Table S2**); only one out of five patients (HGS-24) in our cohort lived shorter than 17 months, and this PDO exhibited the least responsiveness to carboplatin and paclitaxel combination treatment.

A



B

	ORGANOID RESPONSE	CLINICAL RESPONSE				
SAMPLE	AUC	Histopathological* Chemotherapy Response Score	Biochemical* CA-125 normalization	Radiological* RECIST	Progress Free Survival (>6 months)	Overall Survival* (>17 months)
HGS-24	0.88	CRS-1	NO	SD	YES	NO
HGS-13.3	0.75	CRS-1	NO	SD	YES	YES
HGS-6	0.61	CRS-1	NO	PR	NO	YES
HGS-1	0.50	CRS-1	NO	SD	NO	YES
HGS-13.4	0.48	CRS-2	NO	SD	YES	YES
HGS-3.1	0.37	CRS-2	YES	PR	YES	YES
HGS-3.2	0.29	CRS-2	YES	PR	YES	YES

Figure 1: OC PDO drug response correlates with clinical drug response. **(A)** Drug dose-response curves of OC PDOs for carboplatin and paclitaxel combination. Dose response curves are normalized to positive (navitoclax, ABT-263) and negative controls (DMSO). Upper x-axis: carboplatin drug concentrations, lower x-axis: paclitaxel drug concentrations. Each drug concentration was tested twice (technical replicate). Data points and error bars represent the mean and standard deviation of one technical replicate. Non-linear regression analysis: log(inhibitor) vs. response fit. Red=clinically resistant, blue=clinically sensitive. **(B)** Overview of PDO drug response (area under the curve (AUC)) versus all clinical response measures, ordered from low responsive to high responsive based on AUC values. Histopathological tumor response: CRS1=no or minimal response vs CRS2=appreciable response; p=5.821e-05. Biochemical response: no normalization (<35 kU/L) of serum CA-125 during primary treatment vs normalization; p=0.0004. Radiological response: stable disease (SD) vs partial response (PR) according to RECIST criteria; p=0.0092. See also Figure S1D-H. *Statistically significant difference between the clinically sensitive and resistant group according to Wilcoxon signed-rank test corrected for multiple testing (p<0.01).

PDOS EXHIBIT INTERPATIENT DRUG RESPONSE HETEROGENEITY WHICH CORRELATES PARTIALLY WITH THEIR GENETIC MAKEUP

Next, we investigated the response of all PDOs (N=36) to a broader range of drugs and drug combinations (3-17 per PDO, on average 13; **Table S3**), including chemotherapeutics

and targeted drugs. Drugs were selected based on clinical practice or evaluation in clinical trials for ovarian cancer or solid tumors in general. Drug response experiments were performed in technical replicates and replicate AUC values highly correlated (R^2=0.87) (**Figure S3A**). Passage numbers at which PDOs were screened for drug response are provided in **Table S1**. PDOs were classified into a low-responsive subgroup if the drug concentration that reduced viability of >50% of cells (IC50-value) was higher than the concentration achievable in patient plasma (concentration steady state, maximum concentration (C*ss*/C*max*; **Table S3**)[288,289], and a high-responsive subgroup if the IC50-value was lower than the Css/Cmax.

Divergent responses were observed to chemotherapeutic drugs carboplatin (platinum/alkylating agent), paclitaxel (taxane/antimicrotubule agent) and gemcitabine (pyrimidine antagonist) (**Figure 2A-C; Table S3**). A minority of PDOs was classified into the high-responsive subgroup of carboplatin (7/31, 23%) and paclitaxel monotherapy (5/31, 16%), while most PDOs (29/35, 83%) were in the high-responsive subgroup of gemcitabine. Response also correlated with OC histological subtype, all LGS-samples showed low responsiveness to paclitaxel and carboplatin monotherapy (except for response to carboplatin in LGS-3.1), while the high responsiveness was restricted to HG(S)-samples. For certain PDOs, combination treatment with two chemotherapeutic drugs had a greater effect on viability (lower IC50 values) than the drugs' individual effects (**Table S3**) indicating either an additive or synergistic effect of the combined drugs. Our results, for example, showed that carboplatin and paclitaxel treatment alone had minimal effect on LGS-3.1 with an IC50-value of 1.46 and >2.5 log μM, respectively, while the IC50-value of carboplatin and paclitaxel was reduced to 0.56 and -0.34 log μM in the combination treatment.

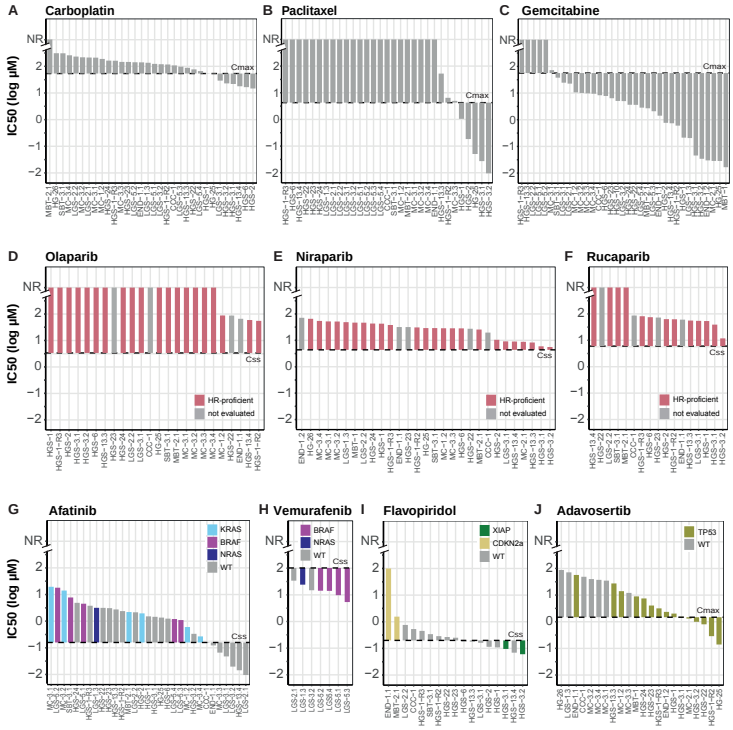


Figure 2. OC PDOs exhibit interpatient heterogeneity in response to chemotherapy and targeted drugs. Waterfall plots with IC50-values (extracted from dose-response curves) of OC PDOs for chemotherapeutics and targeted drugs. The steady state (Css) or maximum (Cmax) *in vivo* plasma concentrations are indicated with the dotted line (Table S3). Bars north of the dotted line represent low-responsive samples, bars south of the dotted line represent high-responsive samples. A-C. Response to chemotherapeutics carboplatin (**A**), paclitaxel (**B**) and gemcitabine (**C**). D-F. Response to PARP-inhibitors olaparib (**D**), niraparib (**E**) and rucaparib (**F**). All PDOs were classified as low responsive to the PARP-inhibitors which correlated with their HR-proficient genetic make-up (no biallelic hit in HR-related genes). Not evaluated = no CHORD evaluation due to missing normal reference. G-J. Response to targeted drugs afatinib (**G**), vemurafenib (**H**), flavopiridol (**I**) and adavosertib (**J**) could partly be explained by genetic aberrations (color-coded) (Table S6, S7). WT=wildtype for the genes mentioned in each panel. NR=IC50-value not reached

The responses to targeted drugs revealed differences and similarities between PDOs which in part correlated to their genetic makeup. For example, all PDOs were classified in the low-responsive subgroup of the PARP-inhibitors olaparib, niraparib, and rucaparib (**Figure 2D-F**), consistent with the absence of biallelic inactivation of *BRCA1* and *BRCA2*, and other genes involved in homologous recombination (e.g. *CHEK2*, *FANCA, PALB2, RAD50, RAD51(B/C/D);* **Table S6 and S7**). Additionally, HR-classifier CHORD classified all samples as HR-proficient based on genome-wide somatic mutation contexts (**Table S7**)[290]. As expected, *BRAF*-, *KRAS*-, and *NRAS*-mutant PDOs were classified in the low-responsive subgroup of pan-HER-inhibitor afatinib (12/25; **Figure 2G**) and high-responsive subgroup of BRAF-inhibitor vemurafenib (5/7; **Figure 2H**). Alterations in *CDKN2A* and *XIAP,* known to affect response to CDK-inhibitor flavo-

piridol, were present in our cohort[291–293]. *CDKN2A* was affected in the two PDOs which were the least responsive to flavopiridol: MBT-2.1 showed loss of both alleles and END-1.1 harbored a nonsense variant (p.(R58*); **Figure 2I**), while copy number loss of *XIAP* was observed in two of the flavopiridol high-responsive PDOs (**Figure 2I**). All *TP53* wildtype PDOs (N=7) were classified into the low-responsive subgroup of WEE1-inhibitor adavosertib, while *TP53* mutants (N=15) were distributed among both the low- and high-responsive PDOs (**Figure 2J**). For the remaining drugs, alpelisib, AZD-8055, MK-2206, and pictilisib, no known genotype and drug response phenotype correlations were observed.

Subsequently, we evaluated for each individual patient how many of the tested monotherapies (3-13 per patient) remained as potential treatments based on an IC50-value smaller than the achievable concentration in patient plasma (C*ss*/C*max*). In case of multiple tumor locations per patient, all test results were considered. A predicted sensitive response (classified in the high-responsive subgroup) to at least one (and maximum five) drug(s) was observed for 88% of patients, except for HGS-1-R3, MC-3 and HG-26 in which all of the 13, seven and three tested monotherapies yielded a predicted resistant response (classified in the low-responsive subgroup) in at least one of their PDOs (**Figure 3; Table S3**).

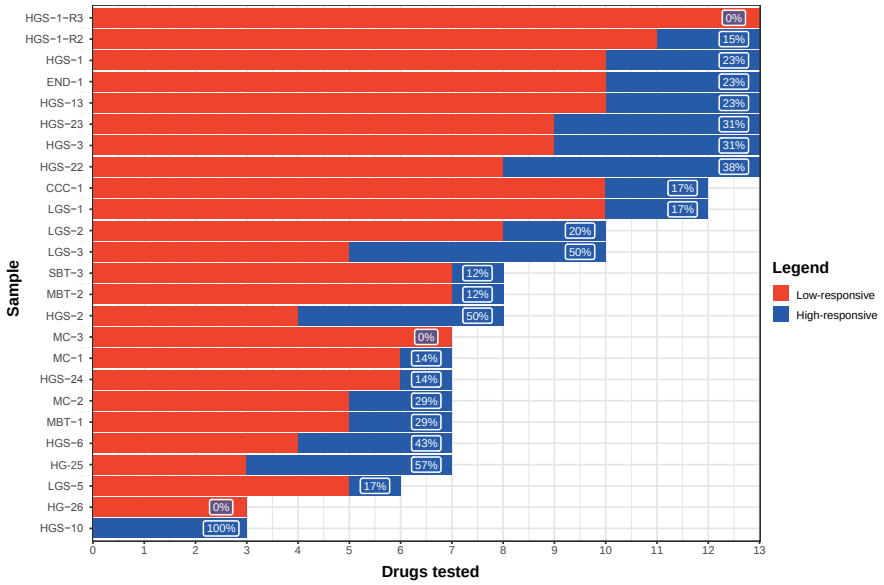


Figure 3: Overview of OC PDO response to single drugs per patient. Overview of the number of monotherapies tested per patient (3-13), classified as low or high responsive, based on the IC50-value relative to the *in vivo* plasma concentration (Css/Cmax, Table S3). For the majority of patients (88%) high responsiveness to at least one tested drug was identified. For patients with organoids derived from multiple tumor locations, results from all tested samples were considered. Red=low responsive, blue=high responsive.

PDOS DERIVED FROM INDIVIDUAL PATIENTS REVEALED INTRAPATIENT DRUG RESPONSE HETEROGENEITY

In addition to assessing interpatient drug response heterogeneity, we examined intrapatient drug response heterogeneity. For seven individual patients, two to four PDOs were derived from distinct cancer lesions at a single time point. For one additional patient, three PDOs were derived at subsequent time points (**Table S1**, heterogeneity comparison). To set a threshold for differential drug response, we first assessed the extent of biological variability. We observed low drug response variability across biological replicates (N=84) with an IC50-value correlation coefficient of R^2=0.82 and mean IC50-fold change of 2.5±1.5 (range=1.0-7.3) (**Figure S3B-C; Table S4**), therefore, a ten-fold change in IC50-value was chosen as a stringent cut-off for differential drug response.

While homogeneous responses were observed to a subset of drugs and drug combinations; carboplatin combined with gemcitabine, adavosertib, or olaparib, carboplatin, olaparib, niraparib, rucaparib, alpelisib, AZD-8055, flavopiridol, pictilisib, and vemurafenib (**Figure S4**), all related PDOs exhibited differential drug response to at least one drug, as defined by a >10-fold change in IC50-value (**Figure 4**). In the seven patients of whom multiple PDOs were derived at the same time point, differential response to mono-treatment was observed 11/36 times (31%). Importantly, in six cases, one of the samples exhibited high responsiveness whereas a related sample exhibited low responsiveness to the tested drugs.

To examine the impact of intratumor genetic heterogeneity on phenotypic heterogeneity, we assessed genetic variants in genes that are known or predicted to interact with drugs according to the drug-gene interaction database resource (DGIdb; **Table S5**)[291]. HGS-13.3, LGS-2.2, LGS-5.2, MC-3.1 and MC-3.2 PDOs were markedly less responsive to the pan-HER-inhibitor afatinib compared to their related PDOs, while differences in response could not be explained by differences in copy number of EGFR/ERBB2(HER2)/ERBB3/ERBB4 (**Figure 4F, Table S7**). Despite meeting the criteria of differential response, all four *BRAF*-mutant LGS-5 PDOs were classified in the low-responsive subgroup of afatinib with an IC50-value above the steady state concentration of -0.8 log μM. The remaining related PDOs with differential response (HGS-13, LGS-2, MC-3) were classified in both the low- and high-responsive subgroup of afatinib. We observed differences in *KRAS* mutation status between the four PDOs derived from a patient with a mucinous OC (MC-3). The two least responsive PDOs (MC-3.1 and MC-3.2) harbored a *KRAS* hotspot mutation (p.G12V), whereas the other low-responsive PDO MC-3.4 harbored two different *KRAS* mutations (p.L19F and p.Q61E, both reported to have an attenuated phenotype compared to hotspot mutations[294,295] and the high-re-

sponsive PDO MC-3.3 was *KRAS* wildtype (WT) (**Table S6**). *KRAS* mutations were independently confirmed with Sanger sequencing (**Figure S3C**).

LGS-5 PDOs also exhibited differential responsiveness to gemcitabine and the MEK-inhibitor cobimetinib (**Figure 4B, 4G**). LGS-5.1 and LGS-5.2 were both in the low-responsive subgroup of gemcitabine, whereas LGS-5.3 and LGS-5.4 were in the high-responsive subgroup. LGS-5.4 was also highly responsive to cobimetinib, while the other LGS-5 PDOs were in the low-responsive subgroup. All LGS-5 PDOs were largely genetically identical, and no variants or copy number changes were identified that explained differential response to these drugs (**Table S6, S7**).

HGS-13.3 PDOs revealed a >10-fold higher IC50-value compared to HGS-13.4 PDOs for gemcitabine, the combined carboplatin and paclitaxel treatment and afatinib (**Figure 4B, 4C, 4F, Table S3**). Consistent with previous findings on the effect of copper-efflux pumps on chemotherapy sensitivity[296–298], copy number losses of *ATP7A* and *ATP7B* were identified in the HGS-13.4 (**Table S7**). Additionally, six other genes previously associated with chemotherapy response (*EIF4EBP1*, *EDNRB*, *NAT2*, *TLE3*, *BRCA2* and *NRG1*)[299–301], exhibited different copy-number states between HGS-13.3 and -13.4 which may also have contributed to the observed differential response to gemcitabine and combined carboplatin and paclitaxel treatment (**Table S7**).

END-1 PDOs, both derived from distinct parts of the tumor lesion in the same ovary, demonstrated differential drug response to gemcitabine, WEE1-inhibitor adavosertib and AKT-inhibitor MK-2206 (**Figure 4B, 4D, 4E**). We identified genetic alterations in *WWOX*, *ERBB2* and *HRAS* that might have contributed to the observed differential response (**Table S6, S7**)[302–306]. However, even though END-1.2 achieved the lowest IC50-values for all three drugs, both END-1.1 and END-1.2 were classified in the high-responsive subgroup of gemcitabine and low-responsive subgroup of MK-2206 and adavosertib (**Figure 4B, 4D, 4E**).

HGS-3.1 and LGS-3.1 displayed drug responses that were very similar to their related PDOs (**Figure 4; Table S3**). In these related PDOs, differential response was only observed for the combined carboplatin and paclitaxel treatment, while drug responses were similar to carboplatin and paclitaxel mono-treatment. Two carboplatin-response associated genes[291], *CLCN6* and *MTHFR*, exhibited copy number loss in the high-responsive PDO LGS-3.1 (**Table S7**). Functional studies have not focused on *CLCN6* and chemotherapy response, but have shown additive effects of *MTHFR*-inhibition and chemotherapeutic drugs[307].

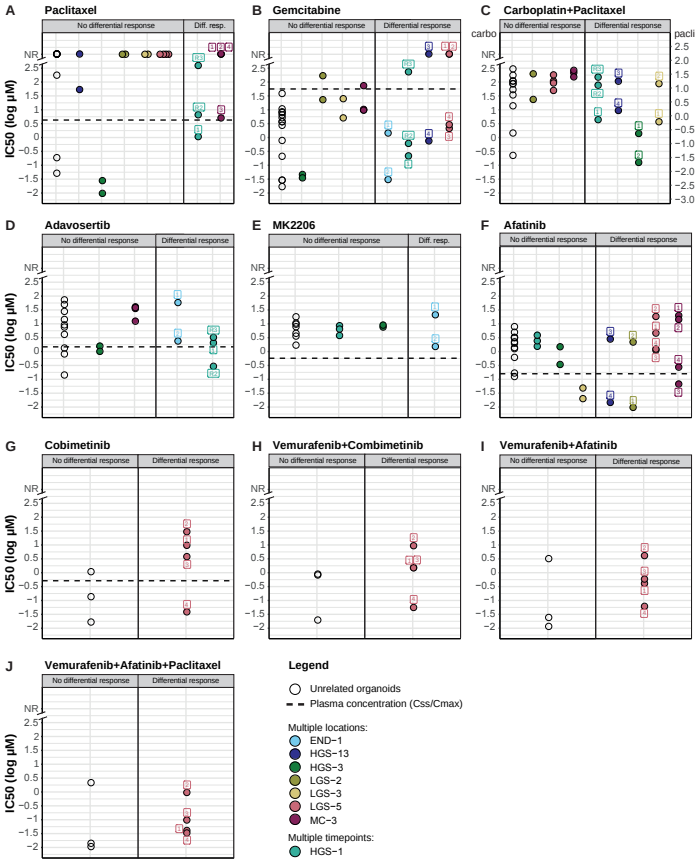


Figure 4: OC PDOs exhibit intrapatient heterogeneity in response to chemotherapy and targeted drugs. IC50-values (extracted from dose-response curves) for drugs that elicit a differential drug response in at least one patient with multiple OC PDOs: paclitaxel (**A**), gemcitabine (**B**), carboplatin+paclitaxel (**C**), adavosertib (**D**), MK2206 (**E**), afatinib (**F**), cobimetinib (**G**), vemurafenib+cobimetinib (**H**), vemurafenib+afatinib (**I**), vemurafenib+afatinib+paclitaxel (**J**). Differential drug response is defined as >10 fold-change in IC50-value within related samples. Left panel: unrelated and related samples without differential response. Right panel: related samples that exhibited differential response. A color code for each patient is shown. The dotted line indicates the steady state (C*ss*) or maximum (C*max*) *in vivo* plasma concentrations for all single drug treatments (Table S3).

Additionally, drug response heterogeneity was examined in a patient from whom PDOs were obtained at multiple time points. HGS-1 was derived from primary chemosensitive disease and HGS-1-R2 and HGS-1-R3 were derived from recurrent chemoresistant disease, and together these PDOs reflected the clinical course of the patient. HGS-1-R2/R3 were less responsive to the mono- and combination treatment of carboplatin and paclitaxel compared to HGS-1 (**Figure 4A, 4C; Table S3**). Although HGS-1-R2 and HGS-1-R3 were derived from ascites collected within a timeframe of one month, differential responsiveness was observed to paclitaxel, gemcitabine and adavosertib (**Figure 4A, 4B, 4D**).

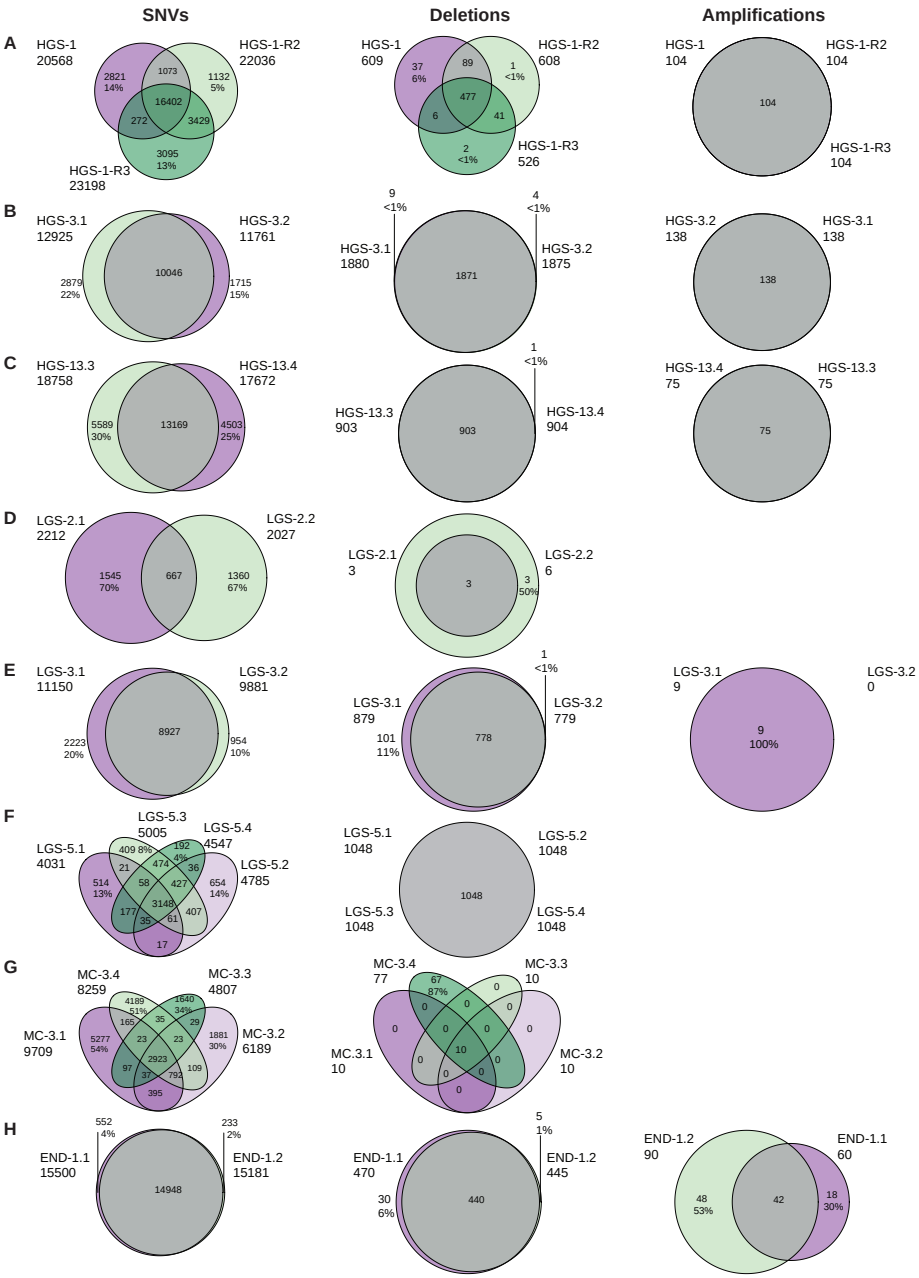


Figure 5: OC PDOs exhibit varying degrees of genome-wide heterogeneity at both the SNV and CNA level. (**A-H**) related OC PDOs with from left to right venn diagrams showing the overlap of all identified SNVs, deletions, and amplifications. In parentheses, the percentage of unique variants in each PDO.

Moreover, we assessed if SVs (including gene fusions) in genes from the DGIdb resource could be linked to differential drug response. In the related PDOs derived from the eight patients that exhibited differential drug response, no SVs were identified that could explain phenotypic heterogeneity. In addition to genetic heterogeneity in genes reported to influence drug response by the DGIdb, related PDOs also exhibited varying degrees of genome-wide heterogeneity at both the SNV and CNA levels (**Figure 5**). The PDOs have been sequenced at slightly different passage numbers, therefore the heterogeneity may also be influenced by tumor content (**Table S7**) and extended passaging, although we have previously shown that PDOs remained similar at the genomic level after extended passaging[278]. The average number of unique SNVs and CNAs in all related PDOs were 24% (2 to 70%) and 17% (0 to 100%), respectively. Considerable genomic heterogeneity at SNV level was observed in LGS-2 and MC-3 (30-70% unique SNVs per PDO), while these PDOs exhibited differential drug response to only one/two drugs. On the other hand, END-1 and LGS-5 had the lowest degree of genomic heterogeneity (2-14% unique SNVs per PDO) and exhibited heterogeneous response to three drugs. In conclusion, we did not observe a direct correlation between genome-wide heterogeneity and differential drug response.

DISCUSSION

We have performed drug screening on 36 PDOs derived from 23 patients comprising all major OC histopathological subtypes. OC PDOs resembled the tumors they were derived from, with an average overlap of 67% of SNVs and similar CNA profiles. PDOs generated at interval debulking recapitulated clinical response to first-line carboplatin and paclitaxel combination treatment for histopathological (p=5.821e-05), biochemical (p=0.0004) and radiological (p=0.0092) outcomes.

Diverse responses to registered drugs for OC were observed among PDOs. Low responsiveness to first-line carboplatin (7/31, 23%) and paclitaxel (5/31, 16%) treatment was observed, compared to high responsiveness in the majority of patients to second-line gemcitabine treatment (29/35, 83%). All PDOs exposed to PARP-inhibition were found to be low responsive, in line with HR-proficiency classification based on WGS data. Response to targeted drugs under clinical investigation could partly be explained by genetic variation; low responsiveness to afatinib (12/25, 48%), low responsiveness to adavosertib (7/17, 41%), high responsiveness to flavopiridol (N=4/17, 24%) and high responsiveness to vemurafenib (5/7, 71%). Importantly, we identified a high responsiveness to at least one tested drug in nearly all patients (22/25, 88%). Since not all PDOs were exposed to the same number of drugs (3-13 monotherapies tested per patient), this is likely an un-

derrepresentation. Finally, intrapatient tumor heterogeneity assessment in seven patients with organoids derived from multiple tumor locations, revealed differential response to at least one drug for all patients, indicating the importance to evaluate multiple tumor locations.

In a systematic approach, we showed that PDO drug response correlated with several clinical response measures. This included histopathological assessment of tumor regression according to a three-tier method (CRS)[285], which is recommended for assessment of response to neoadjuvant therapy[308]. Histopathological grading of tumor regression offers the advantage to study each tumor site separately, as opposed to patient-wide measures of response such as CA-125, RECIST, and survival outcomes. Furthermore, it is a direct measure of chemotherapy response, whereas the survival outcomes may also be influenced by completeness of surgery, co-morbidity and other known prognostic factors. While Bohm *et al.*[285] previously reported that the prognostic significance of the CRS on omental tumor lesions was greater than on primary tumor sites, we applied it to all tumor locations where organoids were derived from. In this study, we present a correlation between CRS and PDO drug response to carboplatin and paclitaxel combination treatment.

In order to bring PDO-based drug response assessment to the clinic, PDO establishment and drug screening needs to be performed within a short timeframe, preferably limited to the diagnosis-treatment interval. In line with previous studies[279,309–311], we demonstrated that PDO establishment and drug screening is feasible within three weeks. Other studies showed a timeframe of one to two weeks. To further validate the predictive value of PDOs, we plan to undertake a prospective trial in which organoids will be derived from both primary and recurrent tumors and tested for response to drugs provided in the clinic, while clinical response is systematically monitored.

Considering intrapatient drug response heterogeneity, derivation of organoids from multiple tumor locations of individual patients, may further improve treatment allocation[312]. Although sequencing studies have shown that OCs display extensive inter- and intratumor heterogeneity on a genetic level[78,283], we could only partially link inter- and intratumor heterogeneous drug responses to genetic heterogeneity. Additionally, some of the CNAs that we identified in genes reported to be related to drug response by the DGIdb might be non-contributive passenger events, given the high frequency of CNAs in high-grade serous OC. Therefore, follow-up studies with increased sample sizes and deeper sequencing are required to decipher drug response associations with the candidate genes identified here, and to discover novel resistance mechanisms. Importantly however, the lack of complete correlation between genetic and functional testing at this

point, stresses the need for functional testing in addition to genetic testing to improve clinical decision making.

The establishment of a larger collection of OC PDOs will provide the opportunity to determine comprehensive, clinically useful genotype-phenotype correlations. When a large collection including drug response data is available, treatment stratification can potentially be performed based on genomic or transcriptomic characteristics of specific PDO subtypes, which could make organoid derivation dispensable in the future[312]. This transition requires an accurate classification of drug-sensitive and -resistant PDOs. Similar to previous studies[288,289], OC PDOs were considered highly responsive if the drug concentration that reduced viability of >50% of cells was lower than the concentration achievable in patient plasma (C*ss*/C*max*). However, the *Css/Cmax* will vary between patients and is not necessarily the concentration that is achieved in the tumor[313,314]. In addition, sometimes patients require dose adjustments due to adverse events which also affect the drug concentration achievable in both plasma and tumor. Therefore, it should be taken into account that tumors predicted to be highly responsive based on the PDO drug response may clinically not respond. Prospective clinical trials comparing clinical to PDO drug response, should be complemented with plasma drug level measurements to further elucidate the relation between *in vitro* and clinical responsiveness.
To conclude, OC PDOs provide a valuable preclinical model system to guide treatment choice in the clinic as it satisfies the following criteria; 1) PDOs genetically resemble the original tumor from which they are derived, 2) PDO drug response often reflects patients' clinical response, and 3) PDO establishment and drug screening can be performed within a short timeframe. Generating and testing PDOs of multiple tumor locations will provide insights in differential drug response as a result of tumor heterogeneity. This information could improve treatment stratification and reduce the development of drug resistance. Complementary PDO drug screening and genomic analysis allows the linkage of genotypes with drug responsiveness patterns to identify candidate biomarkers for drug response.

STAR METHODS

RESOURCE AVAILABILITY

LEAD CONTACT

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ellen Stelloo (estelloo@umcutrecht.nl).

MATERIALS AVAILABILITY

Available OC PDOs are cataloged at www.hub4organoids.eu and can be requested at info@hub4organoids.eu. Distribution of OC PDOs to third parties will have to be authorized by the IRB UMCU at the request of the HUB to ensure compliance with the Dutch 'medical research involving human subjects' act.

DATA AND CODE AVAILABILITY

BAM files of whole-genome sequencing data are made available through controlled access at the European Genome-phenome Archive (EGA) which is hosted at the EBI and the CRG (https://ega-archive.org), under dataset accession number EGA: EGAD00001005707 and EGAD00001004387. Data access requests will be evaluated by the UMCU Department of Genetics Data Access Board (EGAC00001000432) and transferred on completion of a material transfer agreement and authorization by the IRB UMCU at the request of the HUB to ensure compliance with the Dutch 'medical research involving human subjects' act. Additionally, custom code for genomic analyses is available in https://github.com/UMCUGenetics/OvCa_organoids_heterogeneity.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

PATIENT SAMPLES AND CLINICAL DATA COLLECTION

For this study we included women diagnosed with epithelial OC (median age: 65 years). Each patient signed informed consent and was able to withdraw her consent at any time. Tumor samples, ascites and blood samples were gathered between January 2016 and September 2019 at the University Medical Center Utrecht, and Leiden University Medical Center, The Netherlands. Patient data and tissue collection was performed according to the guidelines of the European Network of Research Ethics Committees (EUREC) following European, national and local law. The Institutional Review Board of the UMC Utrecht (IRB UMCU) approved the biobanking protocol: 14-472 HUB-OVI. Clinical data was collected from the patient files. Patient samples were derived at primary disease during primary debulking surgery or interval debulking surgery, or adnex extirpation procedures. Upon recurrence, tissue was collected during (laparoscopic) surgery performed for treatment or diagnostic purposes, or ascites was collected during palliative drainage procedures. No statistical test was used to predetermine sample size.

For the clinical-PDO drug response comparison we selected the samples that met all of the following three conditions: 1) samples were derived at interval debulking surgery; 2) organoid drug response data to carboplatin/paclitaxel was available; 3) patient drug response to carboplatin/paclitaxel was available. For the intrapatient heterogeneity comparison we selected all patients of whom multiple PDOs were derived.

ORGANOID DERIVATION AND CULTURE

Organoids were derived from tumor samples of patients with OC and cultured according to our previously described protocol[278]. Briefly, tumor tissue was cut into small pieces. Two random pieces were separated for fixation in formalin for histopathological analysis and DNA isolation. The remaining tissue was minced, washed with 10 ml advanced DMEM/F12 containing 1x Glutamax, 10 mM HEPES and antibiotics (AdDF+++), collected in a tube, and centrifuged at 300g for 5 minutes. Fragments were allowed to settle under normal gravity for 1 minute, and remaining big tissue pieces were digested in AdDF+++ supplemented with 5 μM RHO/ROCK pathway inhibitor (Y-27632) containing 0.5–1.0 mg/ml collagenase at 37 °C for 0.5–1.0 h. Ascites/pleural effusion samples were centrifuged, and washed with AdDF+++. The cell pellet was allocated fixation in formalin, DNA isolation and organoid derivation. To eliminate erythrocytes, the samples for organoid derivation were incubated with 2 ml red blood cell lysis buffer for 5 min at room temperature followed by an additional wash with 10 ml AdDF+++ and centrifugation at 300g for 5 minutes. Finally, the cells were embedded in BME (Cultrex growth factor reduced BME type 2) on ice and seeded on pre-warmed 24-well suspension culture plates. Following BME polymerization, the cells were overlaid with appropriate organoid culture medium and incubated at 37°C in humidified air containing 5% CO2 (see **Table S1**). PDO names are informative of histological subtype, patient and tumor location. Histological subtype: HGS/LGS=high/low-grade serous adenocarcinoma, HG=high-grade adenocarcinoma, SBT/MBT=serous/mucinous borderline tumor, MC=mucinous adenocarcinoma, CCC=clear cell carcinoma, END=endometrioid carcinoma. The first number indicates the patient, the second number indicates tumor location.

METHOD DETAILS

IN VITRO PDO DRUG RESPONSE TESTING

PDO drug response testing was performed as previously described[278]. In short, PDOs were exposed to drugs in varying concentrations and to controls (DMSO, ABT-263/navitoclax) for 120 hours in 384-well plates, after which ATP levels were measured with the Cell-Titer Glo2.0 assay. All screens were performed in technical replicates. Biological replicates were performed in a subset of PDOs and drugs (**Table S4**) to investigate biological variation. Results were normalized to vehicle (DMSO = 100%) and baseline control (ABT-263/navitoclax 20 μM). Data was analyzed using GraphPad Prism 6. Drug dose-response curves were visualized using linear regression analysis (setting: log(inhibitor) vs. response; least squares (ordinary) fit; top constraint 100%). Concentrations where 50% cell viability (IC50-value) was reached were interpolated. The area under the curve (AUC) was approximated between the lowest and highest concentrations screened

in the actual assay with the trapezoid rule for numerical integration.

CLINICAL DRUG RESPONSE MEASURES

Histopathological response was assessed with the chemotherapy response score (CRS) after three cycles of neoadjuvant chemotherapy, according to the guidelines described by Bohm *et al.*[285]. All available hematoxylin and eosin stained slides of each tumor location from which we established PDOs were assessed for tumor purity. The slide with the most tumor per location was subsequently blinded scored by two certified pathologists (PvD, CV), as CRS-1 (no or minimal pathological response), CRS-2 (appreciable pathological response) or CRS-3 (complete or near-complete pathological response). In case of disagreement consensus was reached. Radiological response was assessed according to the RECIST criteria for solid tumors (version 1.1)[287]. A score for each patient was obtained, from best to worst response: complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD). Biochemical response was measured by assessing response and timing of normalization (<35kU/L) of biomarker cancer antigen 125 (CA-125)[286]. For progression-free survival (PFS) a cut-off of six months was employed; since patients with less than six months PFS are predicted to be resistant to subsequent platinum-treatment. For overall survival, 17 months was taken as a cut-off, based on survival data of a recent cohort of patients (2015-2016) with HGS OC stage IV disease by the Dutch Cancer Registration. Seventeen months after diagnosis, 50% of patients with HGS OC stage IV were still alive.

DNA ISOLATION AND WHOLE-GENOME SEQUENCING

DNA was isolated with the DNeasy Qiagen kit (PDOs and blood samples) and Genomic Tip Qiagen kit (tumor samples), supplemented with RNase treatment. Fresh frozen tumor material obtained through biopsy procedures was processed with the QiaSymphony DSP DNA kit for low input. For DNA library preparation, 500–1,000 ng of DNA was used. Subsequently, whole-genome paired-end sequencing (WGS; 2x 150 bp) was performed on Illumina HiSeq X Ten and NovaSeq 6000 to a median coverage of 31X (range 24-45X).

WGS DATA ANALYSIS

WGS data were processed using the Hartwig Medical Foundation (HMF) somatic mutation workflow. We installed the pipeline (v4.8) locally using GNU Guix with the recipe from https://github.com/UMCUGenetics/guix-additions. Full details and pipeline description are explained in detail by Priestley *et al.*[29] (https://github.com/hartwigmedical/pipeline). Briefly, sequence reads were mapped against human reference genome

GRCh37 with Burrows-Wheeler Alignment (BWA-MEM) (v0.7.5a)[315]. Indel realignment and base recalibration was performed with the Genome Analysis Toolkit (GATK, v3.8.1)[316]. Somatic single nucleotide variants (SNVs) and small insertions and deletions were called with Strelka (v1.0.14)[317]. The functional effect of the somatic SNVs and indels were predicted with SnpEff (v.4.3)[260]. Somatic structural variants (SVs) were called with GRIDSS (v1.8.0)[318]. To assess the SNV overlap between an organoid and a corresponding tumor sample, SNVs that were only detected in either the tumor or the organoid sample of a pair were in a subsequent step called in the corresponding sample (tumor or organoid) when supported by at least one read.

Copy number alterations (CNAs) were called with PURPLE (v2.17)[284]. PURPLE also assesses tumor purity. In case of low tumor purity, a "NO_TUMOR" quality flag was raised by PURPLE, meaning PURPLE failed to find any aneuploidy, and somatic variants were supplied but there were fewer than 300 with observed VAF > 0.1, indicating a high normal cell content (**Table S7**). For tumor samples MC-3.2, MC-3.3, MBT-2.1 and MC-1.2, based on manual verification, a wrong ploidy level was automatically selected by PURPLE. We verified with metaphase spreads analysis on MC-3.2, MBT-2.1 and MC-1.2 PDOs that a ploidy of 2 was most likely for those PDOs and tumor samples. Therefore, and due to the impossibility of manually correcting the ploidy selection in PURPLE, we ran Control-FREEC (v. 11.0)[319] instead on all samples (tumor and PDO) from those patients. Telomeric and centromeric regions were masked for visualization.

For samples CCC-1, END-1.1, END-1.2, HGS-22 and HGS-23 no normal reference sample was available for somatic mutation calling. In these cases, germline SNV calling was performed with GATK[316] and only SNVs with a "HIGH" or "MODERATE" effect as predicted by SnpEff were considered. Similarly, germline SV calling was performed using GRIDSS and SV calls were filtered against the SV Panel of Normals from the HMF analysis pipeline, which can be found in https://resources.hartwigmedicalfoundation.nl. Since PURPLE requires tumor-normal pairs, CNA calling for these five samples was performed individually with Control-FREEC (v. 11.0)[319].

ASSESSMENT OF HOMOLOGOUS RECOMBINATION STATUS

To identify homologous recombination (HR)-deficient samples, *BRCA1* and *BRCA2* as well as other genes in the HR-pathway (*BARD1, BRIP1, CHEK2, FANCA, PALB2, RAD50, RAD51(B/C/D)*) were assessed for biallelic inactivation, incorporating both germline and somatic WGS data. Biallelic inactivation was defined as: a deep deletion (i.e. full loss of both alleles); or Loss-Of-Heterozygosity (LOH) in combination with (i) a known pathogenic/likely pathogenic variant according to ClinVar (https://www.ncbi.

nlm.nih.gov/clinvar/; GRCh37, database date 2018-12-07), or (ii) a frameshift, nonsense or essential splice variant as annotated by SnpEff (http://snpeff.sourceforge.net; v4.1h). Additionally, CHORD (**C**lassifier of **HO**mologous **R**ecombination **D**eficiency, v1.04) was employed to classify PDO samples as HR-proficient or -deficient based on the presence of genome-wide somatic mutation contexts (primarily deletions with flanking microhomology and 1-100kb structural duplications)[290]. Samples without a germline reference sample were excluded from CHORD evaluation.

SELECTION OF GENES FROM THE DGIDB RESOURCE

The Drug Gene Interaction database (DGIdb) was utilized as a resource to obtain a list of genes that have a known interaction with drug response[291]. WGS data of PDOs were checked for SNVs, SVs and CNAs in DGIdb genes, in case differential drug response was observed within related PDOs. Homo-polymer regions were excluded. To identify significantly amplified and deleted genes we applied stringent criteria adopted from Priestley *et al.*[29]. An amplification was defined as [minimum exonic copy number > three times the sample ploidy], while a deletion was defined as [minimum exonic copy number <0.5 times the sample ploidy]. Related samples were regarded genetically heterogeneous on copy number level, if they presented with different copy-number states (amplified vs neutral vs deleted). Furthermore, differential response among related samples was only considered if the ploidy-corrected copy number levels were >10% apart.

QUANTIFICATION AND STATISTICAL ANALYSIS

Descriptive statistics including mean, SD and SEM were conducted with R or GraphPad Prism. The significance level for 95% confidence interval was set to α=0.05. The Pearson correlation test was applied to evaluate the correlation between replicate experiments. The Wilcoxon signed-rank test was applied for the comparison of clinical response groups. The means of two technical replicates of each sample at all measured drug concentrations were compared between clinical response groups (CRS-1 vs -2, RECIST SD vs PR, no CA-125 normalization vs normalization, PFS <6 months vs >=6 months, OS <17 months vs >=17 months). We corrected for multiple testing with the Bonferroni method (alpha = 0.05 / 5 (tests)), resulting in a statistically significant threshold of p=0.01.

LIST OF SUPPLEMENTARY DATA

Figure S1.	Organoid culture and drug response correlation.
Figure S2.	OC PDOs retained genomic features of the original tumor lesions.
Figure S3.	Quality control: drug screening reproducibility and mutation confirmation.
Figure S4.	Drugs that elicit similar drug responses in related OC PDOs.
*Table S1.	Description of study cohort.
*Table S2.	Clinicopathological data for HGS OC PDOs derived at interval debulking surgery.
*Table S3.	IC50-values per drug for OC PDOs related to the maximum (Cmax) or steady state (Css) *in vivo* plasma concentrations.
*Table S4.	IC50-values for all biological replicates.
*Table S5.	Drug response associated genes from the DGIdb resource.
*Table S6.	SNVs in drug response associated genes in OC PDOs.
*Table S7.	CNAs in drug response associated genes in OC PDOs and CHORD classifier scores.

**Table S1-7 are available online at: https://tinyurl.com/Ch3Suppl or scanning the QR code below*

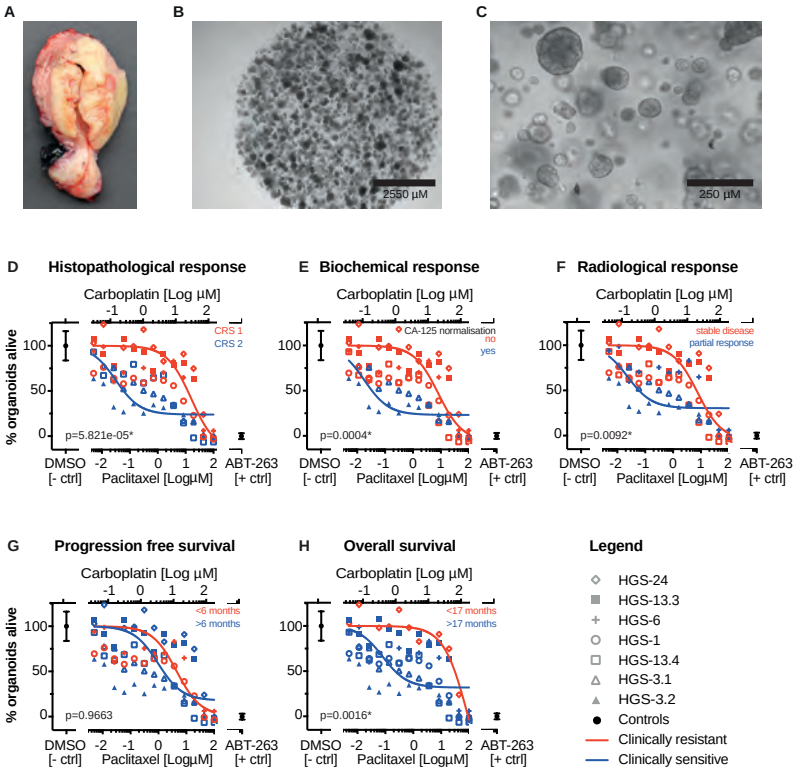


Figure S1: Organoid culture and drug response correlation. Related to Figure 1. (**A**) Macroscopic image of tumor specimen of HG-26 obtained upon recurrence at palliative debulking surgery. Cross-section of the uterus, with exophytic and infiltrating growing tumor, obliterating the uterine cavity. Tumor is perforating deeply into the myometrium. Tumor sample (0.5 cm3) was obtained for organoid culture. For diagnosis of this tumor, the diagnostic department performed an Infinium CytoSNP-850K v1.2 array, which confirmed that the primary seromucinous OC, diagnosed three years earlier, was clonally related to this recurrent high-grade ovarian adenocarcinoma. (**B**-**C**) Representative brightfield images of PDO HG-26 at day 15 prior to rapid drug screening. (**D**-**H**) Correlation of OC PDO drug response with specific measures of clinical drug response, related to Figure 1. Drug dose-response curves of OC PDOs for carboplatin and paclitaxel combination treatment dichotomized into clinical response groups. Each drug combina-tion was tested twice (technical replicate) per OC PDO. Upper x-axis: carboplatin drug concentrations, lower x-axis: paclitaxel drug concentrations. Dose response curves normalized to positive (navitoclax, ABT-263) and negative controls (DMSO). Data points represent the mean of grouped data. Non-linear regression analysis: log(inhibitor) vs. response fit. Red=clinically resistant, blue=clinically sensitive, (**D**) Histopathological tumor response: CRS1=no or minimal response vs CRS2=appreciable response. (**E**) Biochemical response: no normalization (<35 kU/L) of serum CA-125 during primary treatment vs normalization. (**F**) Radiological response: stable disease vs partial response according to RECIST criteria. (**G**) Progression-free survival: <6 months vs >=6 months. (**H**) Overall survival: <17 months vs >=17 months. *Statistically significant difference between the clinically sensitive and resistant group according to Wilcoxon signed-rank test corrected for multiple testing (p<0.01

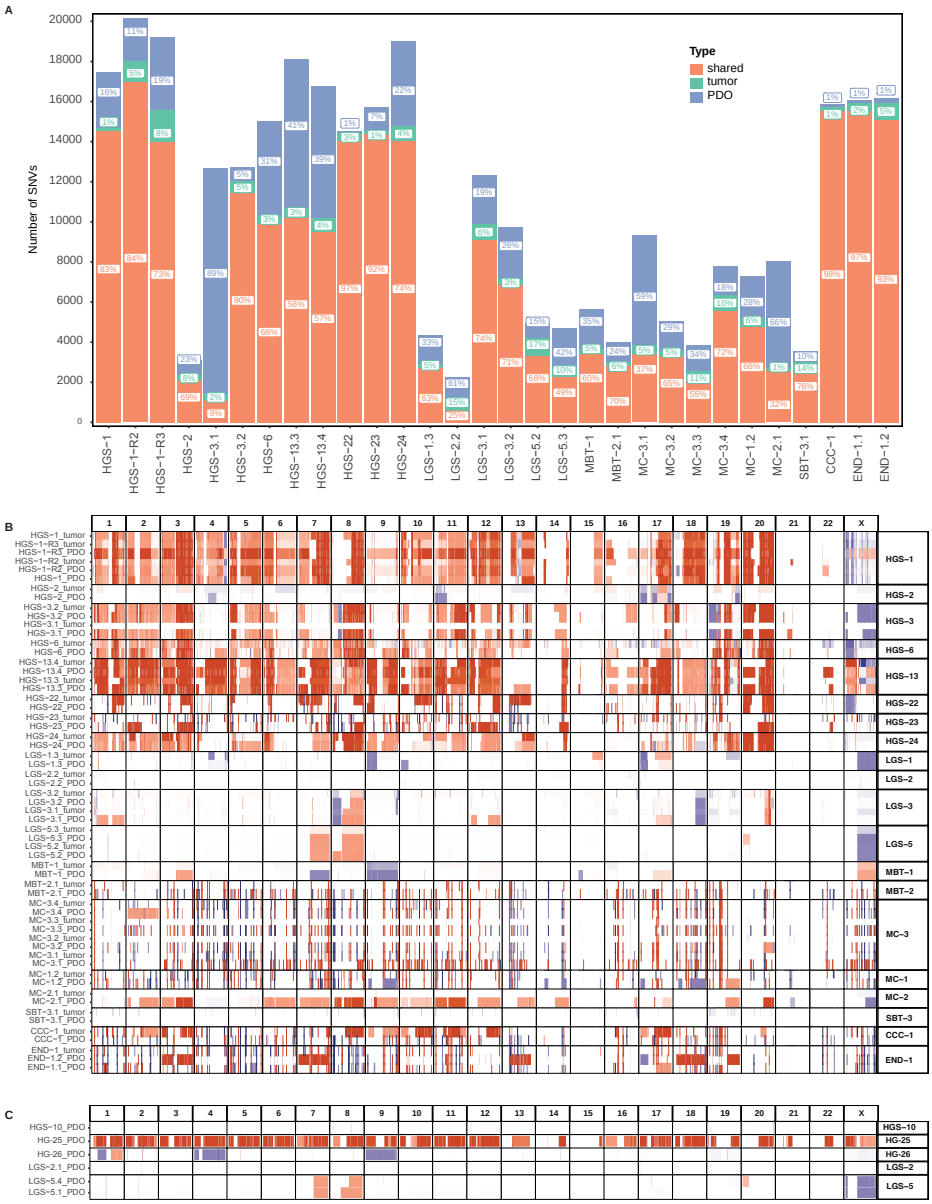


Figure S2: OC PDOs retained genomic features of the original tumor lesions. Related to Figure 5. (**A**) Stacked bar chart showing the number of shared (red) and unique (tumor-green, PDO-blue) SNVs between tumor and PDO pairs. (**B**) Comparison of genome-wide CNAs in tumor and PDO pairs. (**C**) Genome-wide CNAs in PDO-only samples. Copy-number losses are depicted in blue and gains in red.

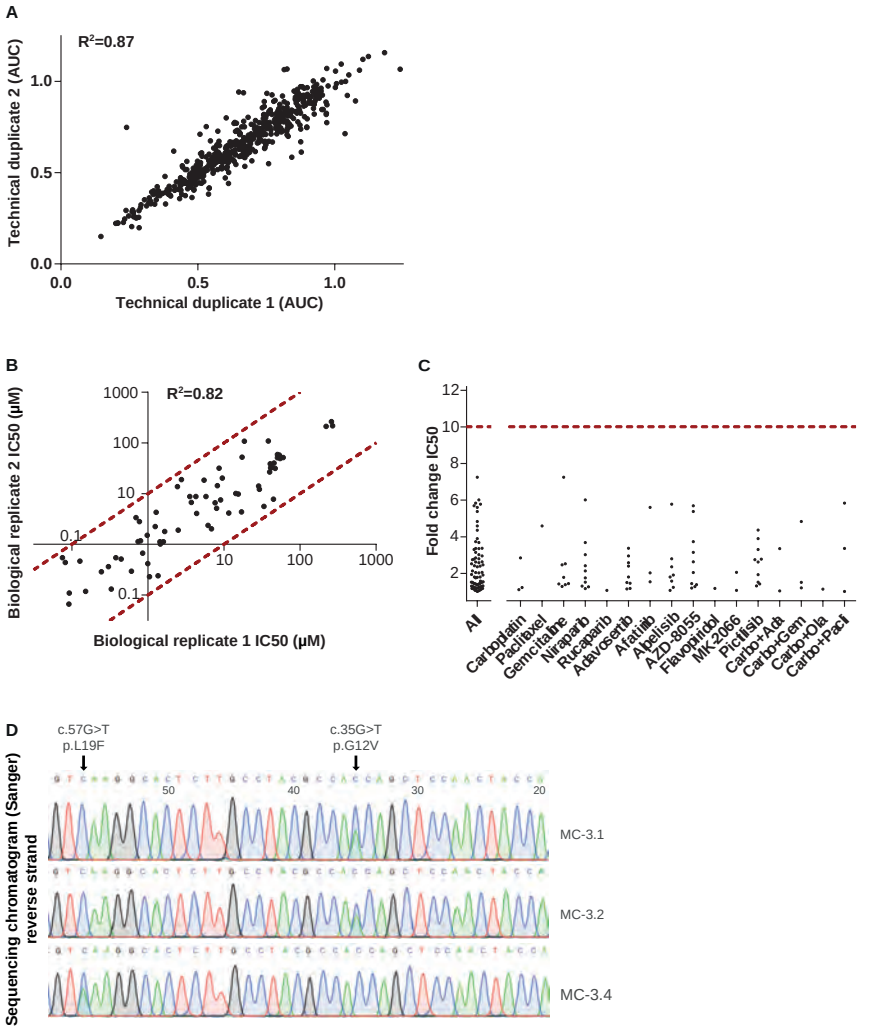


Figure S3: Quality control: drug screening reproducibility and mutation confirmation. Related to Figure 2 and 4. (**A**) Scatterplot of AUC values for all technical replicates of drug screening data. (**B-C**) Reproducibility between PDO biological replicates in terms of drug response. (**B**) Scatterplot of IC50-values for all biological replicates (different passage numbers) for 12 drugs and four drug combination treatments. (**C**) Fold-change in IC50-value between the biological replicates. IC50-values were extracted from the drug dose-response curves. A ten-fold change in IC50-value was chosen as an arbitrary cut-off for differential drug response (red dashed line). (**D**) Confirmation of KRAS muta-tion status by Sanger sequencing in PDOs MC-3. Sequencing chromatogram (reverse strand) for confirmation of KRAS muta-tion p.G12V in MC-3.1 and -3.2 and p.L19F in MC-3.4.

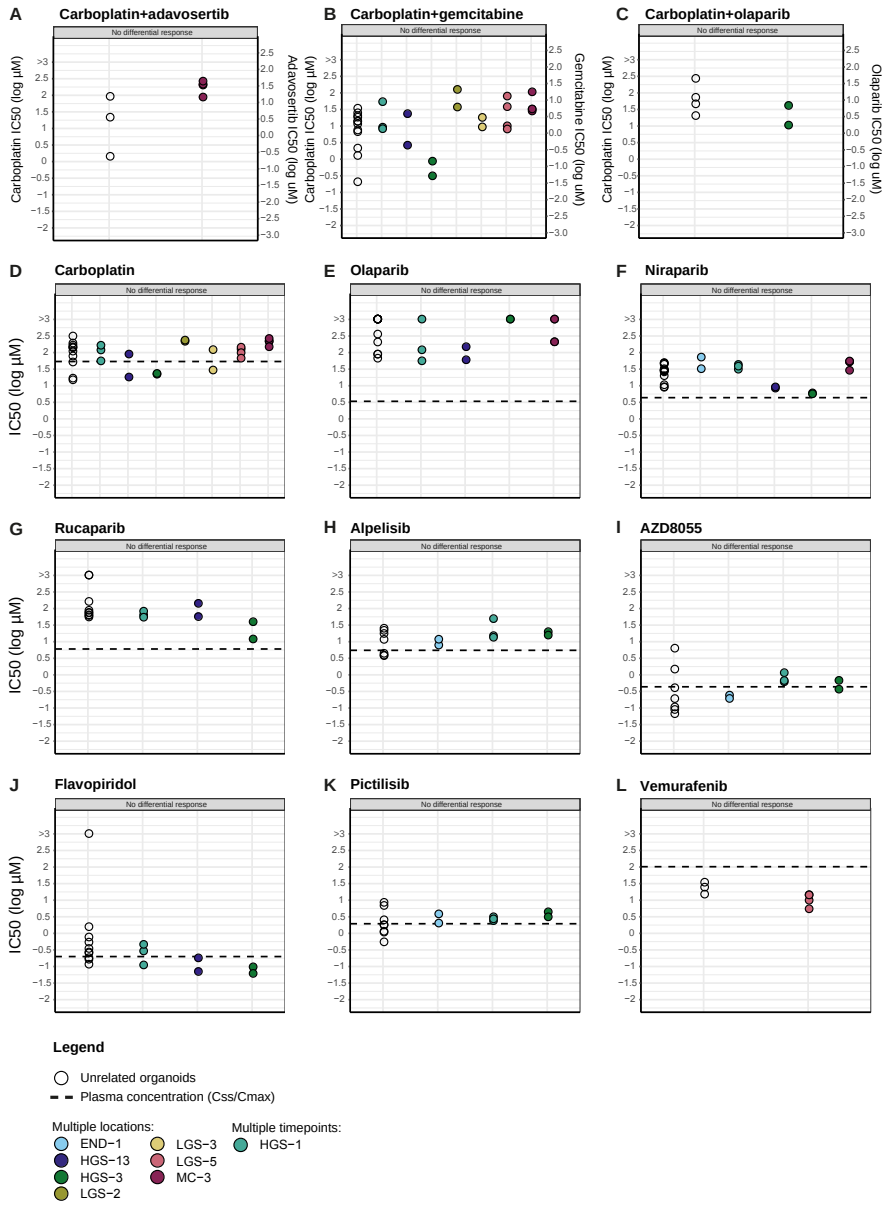


Figure S4: Drugs that elicit similar drug responses in related OC PDOs. Related to figure 4. IC50-values (extracted from dose-response curves) for drugs that elicit similar drug response in all related OC PDOs: carboplatin+adavosertib (**A**), carboplatin+gemcitabine (**B**), carboplatin+olaparib (**C**), carboplatin (**D**), olaparib niraparib (**E**), rucaparib (**F**), alpelisib (**G**), AZD8055 (**H**), flavopiridol (**I**), pictilisib (**J**), vemurafenib (**K**). A color code for each patient is shown. The dotted line indicates the steady state (Css) or maximum (Cmax) *in vivo* plasma concentrations for all single drug treatments (table S3).

ACKNOWLEDGEMENTS

We thank members of the Kloosterman and Cuppen laboratories for helpful discussions; Anne Snelting of the Utrecht Platform for Organoid Technology (U-PORT; UMC Utrecht) for patient inclusion and tissue acquisition; Maaike Vreeswijk and Lise van Wijk (Leiden University Medical Center) for providing ovarian cancer tissues for PDO culturing; Vera Deneer for input on clinical pharmacokinetics; Utrecht Sequencing Facility and Hartwig Medical Foundation for providing sequencing service and data; Hans Bos for acquiring funding; and the Dutch Cancer Registration (IKNL) for providing survival data. The graphical abstract was created with BioRender.com. This work was supported by Gieskes Strijbis Foundation (1816199), and two grants from the Dutch Cancer Society (UU2015-7743, RUG-2017-11352).

AUTHOR CONTRIBUTIONS

Conceptualization: CW, JEVI, RZ, PW, ES
Methodology: CW, JEVI, OK, HC, WK, ES
Software: JEVI, LN
Validation: CW, NH, ES
Formal analysis: CW, JEVI, NH, CV, TJ, PD, ES
Investigation: CW, NH, KL, OK, ES
Resources: CW, JEVI, KL, OK, TJ, LN, RZ, PW, ES
Data curation: CW, JEVI, ES
Writing - original draft preparation: CW, ES
Writing - review and editing: All authors
Visualization: CW, JEVI, ES
Supervision: WK, EC, HS, RZ, PW, ES
Project administration: CW, ES
Funding acquisition: HC, WK, RZ, PW

4

A MULTI-PLATFORM REFERENCE FOR SOMATIC STRUCTURAL VARIATION DETECTION

A multi-platform reference for somatic structural variation detection

Jose Espejo Valle-Inclan[1], Nicolle J.M. Besselink[1], Ewart de Bruijn[2], Daniel L. Cameron[2,3], Jana Ebler[4], Joachim Kutzera[1], Stef van Lieshout[2], Tobias Marschall[4], Marcel Nelen[5], Andy Wing Chun Pang[6], Peter Priestley[2], Ivo Renkens[1], Margaretha G.M. Roemer[7], Markus J. van Roosmalen[1], Aaron M. Wenger[8], Bauke Ylstra[7], Remond J.A. Fijneman[9], Wigard P. Kloosterman[1,$], Edwin Cuppen[1,2,$]

[$]corresponding authors

[1]Center for Molecular Medicine and Oncode Institute, UMC Utrecht, The Netherlands
[2]Hartwig Medical Foundation, Amsterdam, The Netherlands
[3]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
[4]Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Germany
[5]Department of Human Genetics, Radboud UMC, Nijmegen, The Netherlands
[6]Bionano Genomics, San Diego, California, USA
[7]Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Cancer Center Amsterdam, The Netherlands
[8]Pacific Biosciences, Menlo Park, California, USA
[9]Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands

Submitted and available at: bioRxiv 2020.10.15.340497 (2020);
https://doi.org/10.1101/2020.10.15.340497

ABSTRACT

Accurate detection of somatic structural variation (SV) in cancer genomes remains a challenging problem. This is in part due to the lack of high-quality gold standard datasets that enable the benchmarking of experimental approaches and bioinformatic analysis pipelines for comprehensive somatic SV detection. Here, we approached this challenge by genome-wide somatic SV analysis of the paired melanoma and normal lymphoblastoid COLO829 cell lines using four different technologies: Illumina HiSeq, Oxford Nanopore, Pacific Biosciences and 10x Genomics. Based on the evidence from multiple technologies combined with extensive experimental validation, including Bionano optical mapping data and targeted detection of candidate breakpoint junctions, we compiled a comprehensive set of true somatic SVs, comprising all SV types. We demonstrate the utility of this resource by determining the SV detection performance of each technology as a function of tumor purity and sequence depth, highlighting the importance of assessing these parameters in cancer genomics projects and data analysis tool evaluation. The reference truth somatic SV dataset as well as the underlying raw multi-platform sequencing data are freely available and are an important resource for community somatic benchmarking efforts.

INTRODUCTION

Structural genomic variations (SVs) form a major class of somatic genetic variation in cancer genomes[150,320], involving dozens to thousands of somatic SVs with varying size distribution and patterns[150]. While some SVs represent simple deletions, others tend to be complex, involving multiple breakpoints across a relatively short genomic interval. For example, chromothripsis is a form of complex SVs frequently observed in cancer genomes[154,156], resulting from aberrant chromosome segregation or telomere dysfunction[321,322]. Other types of complex SVs involve oncogene amplifications arising from breakage-fusion-bridge cycles[148,150,208]. SVs in cancer genomes may promote cancer development through a variety of mechanisms, such as oncogene activation through gene-fusions, disruption of tumor suppressor genes or by affecting gene regulation[323,324]. Oncogenic fusion genes resulting from somatic SVs form important targets for cancer drugs, and somatic SVs may form neo-antigenic targets for immunotherapies[325], demonstrating the relevance of accurate somatic SV detection for personalized cancer treatment[158,324].

While classical karyotyping and FISH analyses have been instrumental in systematic copy number analyses in tumor samples[158,324], these technologies provide limited resolution or do not allow for comprehensive genome-wide analysis and are thus unable to resolve the complete spectrum of SV events. Most of our knowledge on genome-wide high-resolution SVs in cancer genomes stems from the analysis of short-read whole genome sequencing, which is currently the only scalable and cost-efficient technology for high-resolution genome-wide cancer genome analysis[146,150]. Although short reads are effective for detection of simple SV breakpoints in non-repetitive regions of the genome, the interrogation of complexly rearranged regions or the detection of SV breakpoints in low complexity genomic regions may require other sequencing techniques or targeted approaches[326]. For example, long-insert mate-pair sequencing has proven a valuable strategy for genome-wide mapping of somatic SVs[155,327] and single-cell template strand sequencing enables the detection of copy number variants and copy neutral structural variants[328]. Furthermore, long-read sequencing methods, such as Pacific Biosciences and Oxford Nanopore and synthetic long-read approaches, such as linked-read technology by 10x genomics, provide a promising alternative for the detection of SVs. Initial studies have shown that long-read single-molecule sequencing can greatly improve detection of germline SVs[196,199,200,204]. Similarly, recent work has demonstrated the advantage of long-range sequence information for identification of SVs in cancer genomes, such as cancer gene amplifications and gene fusion events[179,208,211,329].

A major limitation of studies on cancer SVs is the lack of a comprehensive ground truth

genome-wide somatic SV datasets including all types and sizes of somatic structural aberrations. Such truth sets can form a resource for benchmarking sequencing and analysis approaches as well as for evaluating detection problems related to intratumor heterogeneity and tumor purity. Gold reference truth sets have been established for germline SVs[199,213] or somatic single nucleotide variants (SNVs)[214]. However, attempts at benchmarking somatic SVs have only been performed by using *in silico* simulated data[330,331], or mouse data[332].

We addressed this caveat by generating a multiplatform short-read, long-read and linked-read sequencing and optical mapping dataset for the COLO829 melanoma cell line and the paired COLO829BL lymphoblastoid reference cell line. These cell lines have been used before to establish somatic SNV and copy number alteration (CNA) reference sets[214,333,334]. By cross-platform comparison and extensive validation we define a gold reference set of 68 somatic SVs in COLO829. We evaluated the completeness of this validated truth set and demonstrated its use to study the effect of tumor purity and sequencing coverage variation on the accuracy of somatic SV calling. We believe this somatic SV truth set to be of broad value for benchmarking and quality control of large-scale cancer genome sequencing studies, which are currently undertaken in research and the clinic.

RESULTS

MULTI-PLATFORM GENOME-WIDE ANALYSIS OF THE COLO829 TUMOR-NORMAL MELANOMA CELL LINE PAIR

In this study, we aimed at obtaining a comprehensive view on the genome structure of the COLO829 cancer cell line and identify a high-quality set of somatic structural variations, for use as a reference dataset. We cultured COLO829 and the corresponding lymphoblastoid cell line (COLO829BL) according to standard conditions (Materials and Methods). A large batch of cells expanded from one original vial directly obtained from the ATCC cell line repository was used for DNA isolation and subsequent genomic analysis using five different technology platforms: Illumina HiSeq Xten (ILL), Oxford Nanopore Technologies (ONT), Pacific Biosciences (PB), 10x genomics (sequenced on Illumina NovaSeq; 10X), and Bionano Genomics Saphyr optical mapping (BNG) (Materials and Methods).

The sequencing and optical mapping data were analyzed with respect to the reference human genome (GRCh37) using alignment methods specific for each technology (Materials and Methods). From the combined short and long read sequencing data of the

COLO829 sample we obtained a total average base coverage of 235X, while the BNG data generated an additional physical coverage of 218X. For the COLO829BL control cell line a combined average base coverage of 155X and a BNG physical coverage of 220X was reached (**Figure 1A, Supplementary Table 1**). Average physical molecule lengths were 534 bp for ILL paired-end inserts, 11 kb for ONT, 19 kb for PB and 98 kbp for BNG optical maps (**Figure 1B, Supplementary Table 1**).

To assess the consistency of each of the technologies with respect to representation of the sequence content of the COLO829 cancer cell line, we determined the presence of copy number alterations. This revealed a highly similar copy number profile for each of the technologies (**Figure 1C**), with a correlation of copy number calls in the different datasets of 0.87-0.96 (**Supplementary Figure 1A**). Furthermore, we compared our copy number calls with those generated in previous bulk[214] and single cell[335] sequencing of COLO829. The overall CNA landscape of the bulk sequencing and the dominant cluster from single cell sequencing is very similar to the one we obtained (**Supplementary Figure 1B**), with a correlation of 0.99 (bulk) and 0.97 (single cell group A), (**Supplementary Figure 1C**). However, the previously described subclonal single cell clusters (B-D) possess some distinct copy number aberrations that are not observed in our bulk sequencing datasets (i.e. extra copy of chromosome 8 in group D or lack of gain in short arm of chromosome 1), in line with the proposed continuous genomic evolution of cell lines and subculture-specific nature of these events. Finally, classical FISH analysis for six genomic locations of the culture used in our study confirmed the sequencing derived chromosomal copy number states (**Supplementary Figure 3D**).

GENERATION OF A SOMATIC STRUCTURAL VARIATION CONSENSUS TRUTH SET

To build an accurate and comprehensive somatic SV truth set, we used a combinatorial analysis approach involving the four sequencing platforms (ILL, ONT, PB and 10X). Somatic SVs were obtained using state-of-the art SV calling approaches defined for each of the sequencing datasets (Materials and Methods, **Figure 2A**). SV calling parameters were not necessarily optimized for highest precision, but to high sensitivity to not miss out on any real event. As a result, individual candidate call sets for each technology resulted in highly variable lists of predicted somatic SVs, ranging from 92 breakpoint calls in ILL up to 6,412 for ONT, adding up to a total of 8,831 merged candidate somatic SV calls (**Figure 2A**). Only 18 of those somatic SV calls were found by all four sequencing approaches and 125 SV calls were supported by at least two call sets (**Supplementary Figure 2A**).

To make an initial assessment of accuracy, we selected 88 high-confidence SV candidates

for PCR validation based on visual inspection of the mapped reads using IGV. In addition, we randomly selected 296 additional SV candidates for PCR validation. Based on short and long read sequencing of the PCR products, 63 of these breakpoints were labelled as PCR validated (**Supplementary Figure 2B**). Moreover, we decided to perform a separate validation of all 8,831 somatic SV calls from the union of the four SV callsets, using a capture-based enrichment method using multiple probes flanking and overlapping each candidate break-junction (Materials and Methods). Based on the short read sequencing of the enriched products, 114 breakpoints were labelled as capture validated (**Supplementary Figure 2B**). Lastly, we used the 52 BNG somatic SV calls as an additional layer of validation. In total, 137 SV calls were validated by at least one of the methods aforementioned. Additionally, 78 SV calls were not validated but still supported by more than one technology. (**Figure 2A, Supplementary Figure 2C**).

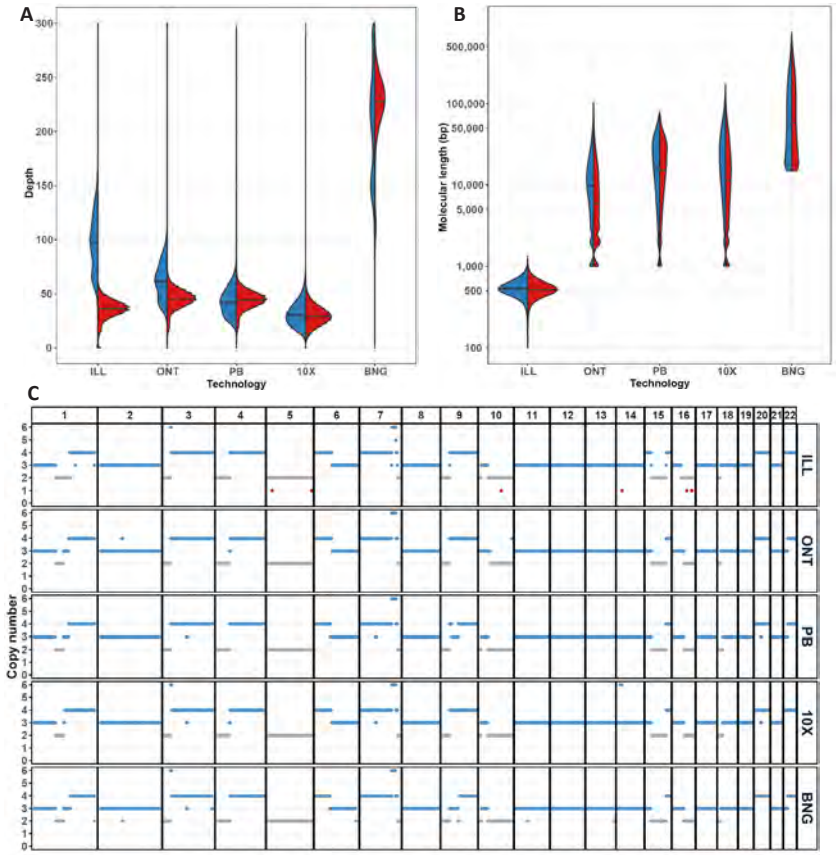


Figure 1: Overview of the COLO829 multi-technology genomic dataset. Sequencing depth (**A**) and log-scaled molecular analysis length (**B**) distributions per technology dataset for COLO829 (blue) and COLO829BL (red). Means are indicated by horizontal black lines. (**C**) Copy number profile of COLO829 calculated independently for each of the datasets.

Next, we manually curated these 215 SV calls that were either validated or supported by multiple technologies. Based on visual inspection of the genomic alignment data from each of the sequencing sets and the validation experiment results, we assessed each SV call individually. We found that 14 calls were actually duplicate calls of the same event (but annotated slightly different by different data analysis pipelines), 48 were real events but also had evidence in the germline control, and another 98 were considered false positive as the supporting or reference data was very noisy at the given genomic location (also in the independent validation data) and may thus reflect the impact of low confidence regions in the reference genome for which unambiguous mapping of sequencing reads is complicated due to simple sequence or repeat content. Taken together, we conclude that 68 of the SV candidates are real somatic events and thus considered our truth set (**Figure 2A, Supplementary Figure 2C, Supplementary Table 2** with all validations and raw calls). To verify the efficacy of our manual curation approach, we randomly selected 179 SV calls that were supported by a single technology and not validated, and therefore left out from the candidate SV curation pipeline, and also evaluated them manually. All these SV calls were either germline events (63, 35%) or false positive due to noisy mapping data (116, 65%) (**Supplementary Figure 2D**).

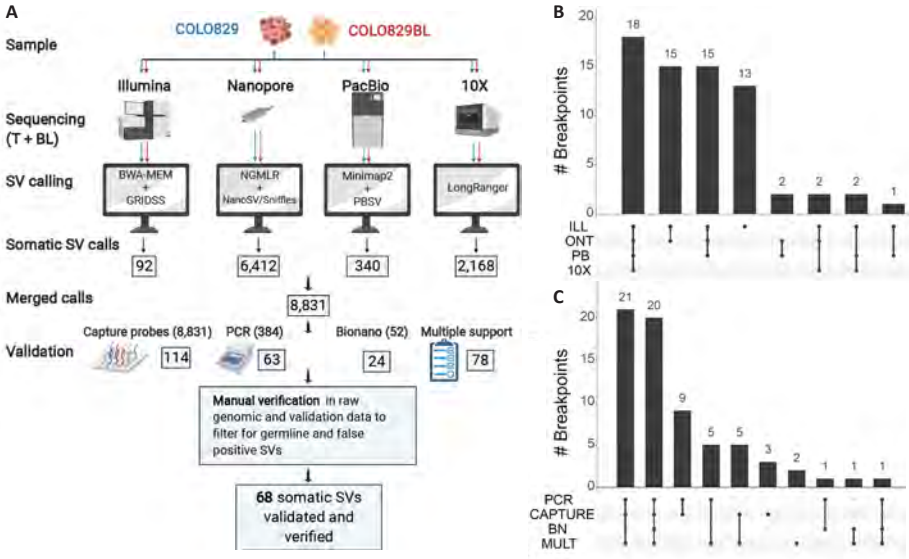


Figure 2: Generation of a validated somatic SV truth set. (**A**) State-of-the-art somatic SV calling pipelines were used independently for each technology dataset. The number of somatic SV candidates identified are indicated in boxes. Overlapping variant calls obtained by the different platforms were merged and independently validated using a combination of targeted enrichment with hybrid capture probes followed by next-gen sequencing, PCR and Bionano genomics. Validated somatic SV candidates and calls supported by more than one dataset were manually curated, leaving a total of 68 somatic SVs in the truth set. Intersections between the 68 somatic SVs in the truth set and the original SV call sets (**B**) and the validation results (**C**) are

shown. ILL = Illumina HiseqX, ONT = Oxford Nanopore, PB = PacBio, 10x = 10x Genomics, BN = Bionano, MULT = support by multiple sequencing platforms.

Of the compiled set of 68 validated somatic SVs in COLO829, 55 (81%) were present in more than two original call sets, including the 18 SVs detected by all technologies (**Figure 2B**). Moreover, most of the SVs were validated at least by capture-based enrichment and by PCR (50, 74%). Additionally, 8 somatic SVs were validated by capture-based enrichment but not by PCR and vice versa, 7 somatic SVs were validated by PCR but not by capture-based enrichment. Of the remaining 3 SVs, one was validated by BNG and 2 were not validated by any targeted assay but are supported by multiple technologies and manually verified by inspection of raw sequencing data from both tumor and normal samples (**Figure 2C**). The resulting somatic SV truth set is presented in **Supplementary Table 3** and freely available as VCF.

CHARACTERIZATION OF THE COLO829 SOMATIC SV TRUTH SET

The somatic SV truth set consists of 38 deletions, 3 insertions, 7 duplications, 7 inversions and 13 translocations (**Figure 3A**). Most of the deletions (24, 61%) are larger than 10kbp, and 7 are smaller than 100bp. There are also three duplications and three inversions larger than 10kbp. Two tumor driver genes are affected by somatic SVs in COLO829 (**Supplementary Table 3**). First, there are two large heterozygous deletions (72 kb and 141 kb) in FHIT, located in the fragile site FRA3B on chromosome 3, which is commonly affected by somatic SVs[150]. Second, there is a homozygous 12 kbp deletion affecting PTEN on chromosome 10.

Frequently, SVs do not occur as simple isolated events but are part of a complex landscape induced in a single event like for example chromothripsis or due to a cascade of events over time like breakage-fusion-bridge cycles. There are also 2 clusters of complex chained somatic SVs that affect 3 or more chromosomes and involve more than 5 breakpoint junctions. Both of them resemble breakage-fusion-bridge events, since they are flanked by foldback inversions and show oscillating copy number profiles[150]. One of them occurred in chromosome 3 and involves four foldback inversions, two of which have templated insertions from chromosomes 10 and 12 and chromosome 6, respectively (**Figure 3C**). The breakpoint and copy number profile of chromosome 3 can be fully explained by 4 cycles of breakage-fusion-bridge followed by chromatid duplication through a whole genome doubling event. Initiated by replication of unrepaired double-stranded break, the unstable chromosome 3 (due to the presence of two centromeres in a single chromatid) underwent a further 3 more rounds of BFB with a fragment of chromosome 6 inserted prior to the third doubling cycle, fragments of chromosomes 10 and 12 inserted immediately after the fourth doubling cycle, and a stable state achieved

after the final breakage through repair to one of the centromeres (**Supplementary Movie**). The other breakage-fusion-bridge event occurred on chromosome 15 and includes templated insertions from chromosomes 6 and 20 (**Figure 3D**). The donor locations of these templated insertions are not affected by SV events.

To evaluate the completeness of the somatic SV truth set, we compared it with the somatic CNA calls, since each CNA should have SV breakpoints or telomeres at either end. We found 43 total CNA breakpoints that are not telomeric ends of chromosomes. Of these, 26 (60%) are concurrent with an SV breakpoint. We evaluated the rest of the CNAs in the raw genomic data (**Supplementary table 4**). Six more copy number breakpoints (14%) are present in the germline, flanking heterozygous germline CNA events that are homozygous in the tumor through a somatic loss of the other allele. The SV break-junctions of these CNAs are germline and therefore not part of the truth set. Finally, there are 11 somatic CNA breakpoints (26%) not concurrent with an SV breakpoint. Five of these missing CNA breakpoints are located in a centromeric region (chromosomes 1, 4, 6, 14 and 16) and are likely due to a missing somatic SV involving the centromere, which are typically hard to fully resolve due to their repetitive nature. For another 2 missing CNA breakpoints (chromosome 3 and chromosome 9) breakends can be found in the raw ILL dataset, meaning an SV breakpoint was found but the SV junction partner could not be unequivocally determined. GRIDSS2 annotation did reveal that the chromosome 3 single break does map to one of the centromeres. Four more missing CNA breakpoints flank two supposed deletions in chromosome 1, but no SV call in these locations can be found for either COLO829 or COLO829BL in any of the datasets. Manual inspection of the raw data for these CNAs (**Supplementary Figure 3A, 3B**) indicates that these CNAs may actually reflect heterozygous germline events followed by LOH as witnessed by the lower read coverage in the COLO829BL as compared to the flanking segments. Furthermore, one CNA involves a LINE-rich region while the other overlaps with a segmental duplication.

Next, we compared our somatic SV truth set to the somatic SV calls presented by Arora *et al.* They provide two different somatic SV callsets, one generated by the HiSeq platform with 77 somatic SV calls and the other by the NovaSeq platform with 75 somatic SV calls. Since these were provided based on GRCh38 genomic coordinates, we lifted our somatic SV coordinates over to GRCh38. We found that 58 (75.34%) and 59 (78.6%) of the somatic SV calls for the HiSeq and the NextSeq callsets, respectively, overlapped with our somatic SV truth set on both sides of the SV (**Supplementary Figure 3**). We manually inspected the 20 non-overlapping somatic SV calls from the Arora et al dataset in our raw ILL, ONT and PB data (**Supplementary Table 5**). In the long-read raw data (ONT and PB) only 3 out of the 20 have some support (maximum 3 reads). In the ILL raw data,

9 out of the 20 have limited evidence, with only one or a few supporting reads. Only 4 of these 9 SV calls passed bioinformatic calling criteria in our original ILL somatic SV calls, but none of these were called by any other technology or independently validated by more sensitive PCR or targeted capture and deep-sequencing. Therefore we consider these candidates as technology-specific noise and were discarded from our truth set, although we can formally not exclude that these are real variants that are present at very low frequency (<1% in the sample). Finally, 13 SVs are present in our truth set and not in the Arora *et al.* data set. All were detected by at least two different sequencing techniques and independently validated.

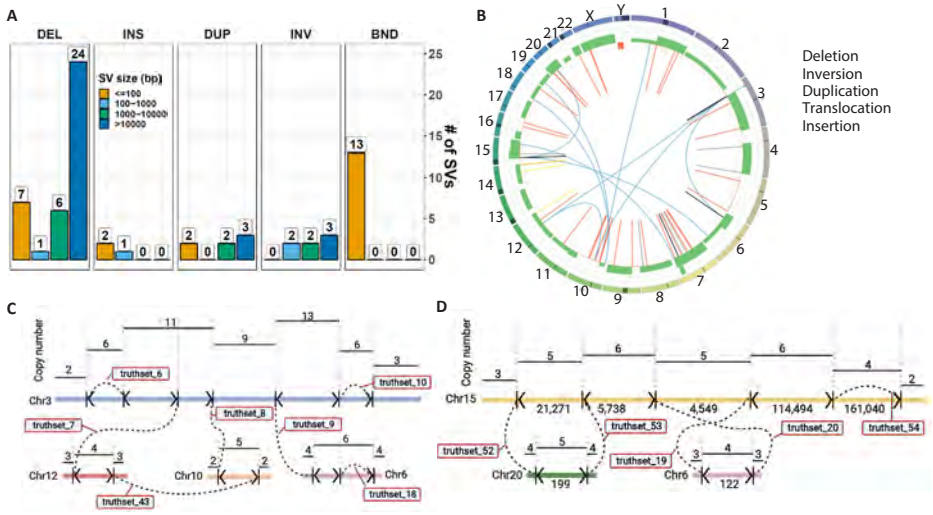


Figure 3: Characterization of the somatic SV truth set. **(A)** Distribution of different types of SVs in the COLO829 truth set, divided in size bins. Translocations (BND) are assigned a size of 0 bp. **(B)** Correlation between CNAs and somatic SVs in the COLO829 truth set. The circos plot shows copy number gains (green) and losses (red) and somatic SVs. Each copy number change is expected to be flanked by an SV event. Two complex breakage-fusion-bridge events are present in COLO829. The first one **(C)** occurs in chromosome 3 (blue), with templated insertions from chromosomes 6 (pink), 10 (green) and 12 (red) (see also Supplementary Movie for an animation of the proposed mechanism shaping this event). The second one **(D)** occurs in chromosome 15, with templated insertions from chromosomes 6 (pink) and 20 (orange). Breakpoints are indicated by vertical lines with arrowheads showing breakpoint orientations. Dashed lines indicate junctions between two breakpoints. Break-junctions are labelled with truth set SV ID number (Supplementary Table 3).

EFFECT OF TUMOR PURITY AND SEQUENCING DEPTH ON SOMATIC SV CALLING

To demonstrate the utility of the COLO829 somatic SV truth set, we evaluated the effect of tumor purity, which is highly variable amongst clinical samples, on SV calling. We used the available raw datasets and simulated tumor purities of 75% (TP75), 50% (TP50), 25% (TP25), 20% (TP20), and 10% (TP10) by random *in silico* mixing of the genomic data from COLO829 and COLO829BL for ILL, ONT and PB, respectively. We performed SV calling independently on each of these mixed sets and on the original tumor file (100% purity, TP100) and the normal file (0% purity, TP0). We then calculated the recall (percentage of truth set found) and precision (percentage of calls that belong to truth set). With the standard settings used, somatic SV recall and precision were found to be highly dependent on tumor purity for all three technologies. At TP75 and TP100, recall is the highest, with >94% for ILL, >67% for ONT and >65% for PB. With TP50, the recall slightly decreases to 90%, 52% and 61% for ILL, ONT and PB, respectively. For purities lower than TP50, the recall decreases further to <76%, <22% and <48% for ILL, ONT and PB, respectively. Precision follows a similar trend in the case of ILL, with precisions >78% for purities larger than TP50, and a drop to 63% in TP25. In the case of ONT and PB, the higher number of false positives impact severly on the precision rates, potentially reflecting maturity level of platform-specific tools for somatic SV detection in tumor-normal paired samples, but also presenting opportunities for further analysis parameter and tool optimisation.

Sequencing depth is another essential parameter to consider in tumor sequencing projects as it represents a trade-off decision between variant detection sensitivity and costs. To investigate the effect of sequencing depth in combination with tumor purity in somatic SV detection, we took one of the triplicates from each of the simulated ILL tumor purities (98x coverage) and subsampled them to 50x, 30x, 10x, 5x and 1x depths. We again performed somatic SV calling using the same standard pipeline on each of these simulated sets and calculated recall and precision. We observed that for depths of 50x and 98x and tumor purities over 50% recall was over 82%. In the case of 98x, even at TP20 a recall of 71% could be obtained, whereas for 50x at TP25 the recall decreased to 42%. For 30x sequencing depth, at TP100 recall was 84%, but at TP50 there was a decrease to 54% and at TP25 further to 10%. For lower coverages, recall was low. Surprisingly, depths of 30x and 50x had a higher precision at all tumor purities than 98x, with precision around 95% over TP50, compared to approximately 70% for 98x. While this could in theory be explained by the presence of subclonal SVs that are not included in the reference truth set but become detectable at higher sequencing depth, this might also be caused by stochastic effects due to increased measurement noise at higher sequencing depth which increases the number of false positive and therefore reduces precision (although recall is not affected). Further optimization of analysis tools and settings and deeper sequencing may resolve these issues.

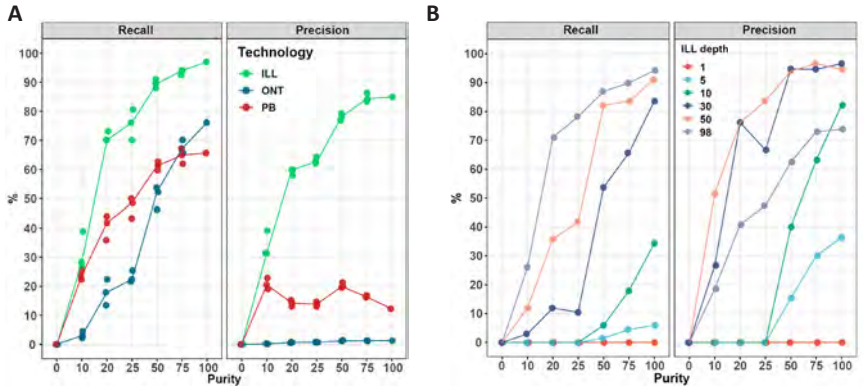


Figure 4: Recall and precision of somatic SV calling as function of tumour purity and sequencing depth effect. Different tumor purities (0, 10, 20, 25, 50, 75 and 100 %) were simulated by mixing data from COLO829 and COLO829BL for the ILL, ONT, and PB datasets. **(A)** Somatic SV calling was performed independently for each purity subset and recall (left) and precision (right) were calculated against the COLO829 somatic SV truth set. Lines represent the median of independent triplicate measurements. **(B)** For each tumor purity subset in the ILL dataset, different sequencing depths (1, 5, 10, 30, 50 and 98x) were sampled. Somatic SV calling was performed independently for each sequencing depth and tumor purity subset and recall (left) and precision (right) were calculated against the COLO829 somatic SV truth set.

DISCUSSION

We produced a validated somatic SV truth set by building upon the strengths of different sequencing technologies. Bioinformatic integration of results and large-scale independent validation strategies turned out to be a powerful approach for reducing the large number of candidate events obtained. Manual curation and inspection of raw sequencing data was however essential to exclude sequencing, mapping artefacts and remaining germline events. These somatic false positives are thus germline false negatives and were likely included in the initial somatic SV calls due to the lower sequencing analysis depths for the control sample as compared to the tumor (typically 3-fold lower) in combination with specific local genomic characteristics (e.g. lower average coverage due to for example local GC content or involving low complexity sequences)[336].

While reconstruction of the derived chromosomal tumor genome topology based on the 68 truth set somatic SVs results in an overall stable genomic configuration for most derived chromatids harboring a single centromere and two telomeres, some breakpoint junctions are still clearly missing. This is corroborated by the fact that not for all copy number alterations breakpoint junctions were identified at either end. Our results indicate that these missing events typically involve centromeric regions that are not directly accessible by any current sequencing technology. Annotation data provided by the

GRIDSS2 SV caller[337]. suggests a junction between a single break-end in chromosome 3 and the centromere in chromosome 1, which shows a copy number change. This can probably not be resolved directly due to the repeated nature of the centromeric region. When excluding the missing events that likely involve centromeres, there are 2 copy number aberrations that remain unexplained by the truth set, providing room for further improvement based on the existing or to be generated data.

Although this study was not designed to compare performance of sequencing platforms or data analysis pipelines, some interesting observations can be made. First, there is clear complementarity between the various platforms for the comprehensive identification of all real events. However, bioinformatic pipelines for somatic SV detection are still clearly in different stages for the different platforms with the most commonly used Illumina-based approaches yielding lowest numbers of false positives. We believe future tool optimisation for somatic SV calling, assisted by gold reference truth sets as well as the development of platform-specific germline and artefact filtering data sets ('pools of normals') based on large numbers of samples, will effectively address this challenge. Second, data analysis pipelines yield different annotations for the same event. This calls for further standardisation of variant annotation and nomenclature, although some observed differences are intrinsic to the use of short and long-read technologies. For example, a long templated insertion may be called as two independent translocations by short-read SV callers, while long read-based technology would detect this readily as an insertion. Third, despite previous studies showing the added value of long reads for SV detection for germline events, our somatic SV truth set is resolved almost in its entirety with the ILL short read dataset. While this may in part be due to the more advanced somatic SV calling pipelines developed for short-read data, this observation may also be explained by fundamental differences between germline and somatic SVs, where the latter are much more randomly distributed throughout the genome than inherited germline events. Germline variants more often involve complex or repetitive regions of the genome which might reflect mechanistic differences like for example the more frequent involvement of non-allelic homologous recombination, or be due to differences in selective pressure. As a consequence, somatic events may thus on average be more effectively detected.
The COLO829 cell line has the advantage that it is, in contrast to real tumor samples, a renewable source that can be used for assessing the impact of future platform developments or the performance of completely new technologies for somatic mutation detection. Although the COLO829 cell line is representative for structural variation as observed in cancer, including small and large copy number alterations (including aneuploidies) and both simple and complex SV events, it is not necessarily representative in all aspects for real tumor samples. First, tumor samples do typically not consist of tumor cells only but are a mix of tumor and normal cells (e.g. stromal cells and infiltrating immune cells). We

show that the raw data obtained in this study can be used effectively to mimic variable tumor purity and that the truth set is instrumental for assessing the performance of the bioinformatic data analysis tools at variable tumor purity. As expected, our results show that both recall and precision heavily depend on tumor purity for all platforms. Secondly, tumors evolve continuously and are typically genetically heterogeneous, especially primary tumors, involving potentially subclonal SV events. While the COLO829 cell line has been shown to be genetically heterogeneous and evolving over time and thus could in principle capture this tumor feature properly, this variation is dynamic and might be variable between cell line isolates as already demonstrated by the various studies on this cell line[334,335] and thus limit the utility of a single defined truth set obtained as presented here. Finally, tumors are in general very heterogeneous both within the context of a specific tumor type, but especially between tumor types. For example, microsatellite instable (MSI) tumors show a high number of small indels[338], homologous recombination deficient (HRD) tumors present many deletions with microhomology and large duplications[290] and paediatric haematological cancers cancers usually show very low mutational load but enhanced levels of somatic SVs, although often involving specific but complex genomic loci (e.g. the IgH locus)[31,339]. The specificity for capturing such heterogeneity effectively or the impact of specific genomic events that may co-occur in a given tumor sample, like for example whole genome duplication or chromothripsis, on overall performance of a specific sequencing technique or data analysis tool is of course not captured in a single cell line and requires the development of complementary datasets. Analysing additional cancer cell lines with matching normal cell lines provide an attractive route towards this goal as these represent in principle an endless source of genomic material for benchmarking of future DNA analysis technologies, but also for quality monitoring in routine production labs under accreditation. However, availability of suited cell lines that represent the full genetic diversity of cancer is a clear limitation. Ideally, one would thus resort to thoroughly analysed real tumor samples, even though in practice availability of sufficient material for multi-lab and multi-technology analyses can be problematic and sharing and reusing of patient material and data may require complex consenting and legal procedures.

Taken together, we believe the SV truth set described here as well as the underlying raw data, are a valuable resource for benchmarking and fine-tuning analysis settings of somatic SV calling tools, but the data may also be used for the development of novel analysis tools, for example phasing of structural variants. All analysis results and raw data are publicly available to enable such applications without access restrictions (ENA accession number: PRJEB27698 and an overview of the available data and specific access link can be found at **Supplementary Table 6**). We demonstrate this utility by analysing the impact of tumor purity and sequencing depth on SV recall and precision for different

technologies, thereby providing valuable insights in the potential impact of technology platform choice and experimental design in relation to diagnostic accuracy and overall costs. Furthermore, these results highlight the need of benchmarking somatic SV detection methods at different tumor purities and sequencing depths rather than under a single fixed condition, since these parameters are highly variable within and between cohorts and can result in strong performance variation.

MATERIALS AND METHODS

Sample source: COLO829 (ATCC® CRL-1974™) and COLO829BL (ATCC® CRL-1980™) cell lines were obtained from ATCC in September 2017. A single batch of cells was thawed and cells were expanded and grown according to standard procedures as recommended by ATCC. Cell pellets were split for technology-specific DNA isolation at 33 days (COLO829 & COLO829BL for the ILL and ONT datasets), 35 days (COLO829 for the PB, 10X and BNG datasets) and 23 days (COLO829BL for the PB, 10X and BNG datasets).

GENOMIC ANALYSES PER TECHNOLOGY

Illumina: COLO829 and COLO829BL libraries were prepped with Truseq Nano reagent kit and sequenced on the HiSeq X Ten platform using standard settings and reagent kits (chemistry version V2.5). Reads were mapped to GRCh37 with BWA mem (version 0.7.5)[340], followed by indel realignment with GATK (v3.4-46)[341]. SVs were called jointly for COLO829 and COLO829BL with GRIDSS (v2.0.1)[337]. Somatics SVs were filtered with the GRIDSS somatic SV filtering script (https://github.com/PapenfussLab/gridss/blob/master/scripts/gridss_somatic_filter.R).

Nanopore: COLO829 and COLO829BL libraries were sequenced on the MinION and GridION platforms using R9.4 flow cells. Reads were mapped to GRCh37 with NGMLR (v0.2.6, default parameters)[342] with default parameters. SV calling was performed with both NanoSV (v. 1.2.2, default parameters)[204] and Sniffles (v1.0.9, parameters "*--report_BND --genotype*")[342] for COLO829 and COLO829BL separately. All SV calls for both NanoSV and Sniffles were merged with SURVIVOR (v1.0.6)[343] with a distance of 200 bp and calls with evidence in COLO829BL for NanoSV or Sniffles were discarded.

PacBio: COLO829 and COLO829BL libraries were sequenced on the Sequel System with the 5.0 chemistry (binding kit 101-365-900; sequencing kit 101-309-500). Reads were mapped to GRCh37 with minimap2 (v2.11-r797)[344]. SVs were called jointly for COLO829 and COLO829BL with pbsv (v2.0.1, https://github.com/pacificbiosciences/

pbsv/) using default parameters. Somatic SV calls were filtered by removing any call with a supporting read in COLO829BL.

10X: COLO829 and COLO829BL 10x genomics libraries were prepared on the Chromium platform and sequenced on the NovaSeq platform (chemistry version V1). Reads were analyzed with the LongRanger WGS pipeline (v2.2.2) separately for COLO829 (somatic mode) and COLO829BL (default parameters). SV calls for COLO829 and COLO829BL were merged with SURVIVOR (v. 1.0.6)[343] with an overlap distance of 200 bp and SV calls with evidence in COLO829BL were discarded.

Bionano: DNA for COLO829 and COLO829BL was labelled using the Bionano Direct Label and Stain (DLS) kit. The labelled DNA was linearized in a Saphyr chip and imaging was performed on the Saphyr instrument. SV calling was performed on the Bionano Access platform. For each sample, 1.5 million cultured cells were used to purify ultra-high molecular weight DNA using the SP Blood & Cell Culture DNA Isolation Kit following manufacturer instructions (Bionano genomics, San Diego USA). Briefly, after counting, white blood cells were pelleted (2200g for 2mn) and treated with LBB lysis buffer and proteinase K to release genomic DNA (gDNA). After inactivation of proteinase K by PMSF treatment, genomic DNA was bound to a paramagnetic disk, washed and eluted in an appropriate buffer. Ultra-High molecular weight DNA was left to homogenize at room temperature overnight. The next day, DNA molecules were labeled using the DLS (Direct Label and Stain) DNA Labeling Kit (Bionano genomics, San Diego USA). Seven hundred and fifty nanograms of gDNA were labelled in presence of Direct Label Enzyme (DLE-1) and DL-green fluorophores. After clean-up of the excess of DL-Green fluorophores and rapid digestion of the remaining DLE-1 enzyme by proteinase K, DNA backbone was counterstained overnight before quantitation and visualization on a Saphyr instrument. A volume of 8.5 microliter of labelled gDNA solution of concentration between 4 and 12ng/µl was loaded on the Saphyr chip and scanned on the Saphyr instrument (Bionano genomics, San Diego USA). A total of 1.6 Tb and 1.5 Tb of data was collected for the cancer and blood sample, respectively.

De novo assembly Pipeline and Copy number variants calling were performed and against the Genome Reference Consortium Human Build 37 (GRCh37) HG19 human genome assembly (RefAligner version 7520). Events detected by the de novo assembly pipeline were subsequently compared against the matched blood control, and those that are absent in the assembly or the molecules of the control were considered as somatic variants.

Consolidation of SV calls: Somatic SV calls for each dataset (ILL, ONT, PB and 10X)

were merged using SURVIVOR (v. 1.0.6) [343] with an overlap distance of 200bp.

DEPTH AND MOLECULAR LENGTH CALCULATIONS

Average base depth and depth distribution for ILL, ONT, PB and 10X was calculated based on 1,000,000 random positions on the genome with Sambamba (v0.6.5)[345]. Average base depth for BNG was calculated based on the same 1,000,000 random positions using Bedtools (v2.25.0)[346].
Average molecular length and length distribution was calculated based on insert size for ILL, read length for ONT and PB, on synthetic molecular length based on the MI tag for 10X, on optical map length for BNG. For ILL, average insert size was calculated using Picard (v1.141, http://broadinstitute.github.io/picard).

COPY NUMBER ANALYSIS

CNA calling was performed on the ILL dataset with BIC-SEQ2 (v0.7.2)[347]. For the remaining datasets, BAM and optical map (xmap) files were converted to BED format using Bedtools (v2.25.0)[346] and CNA calling was performed with Ginkgo[348]. CNA calls from the different datasets were merged using 1MB bins to calculate Pearson's correlation between datasets and for plotting.

VALIDATIONS

Capture: For each break-junction of the merged somatic SV calls 2 capture probes of 100 bp in length were designed, one at either side of the breakpoint, with a maximum distance of 100bp from the breakpoint at GC percentage as close as possible to 50%, for a total of 18148 custom probes. These custom capture probes were then ordered from Twist Biosciences. Then, libraries for COLO829 and COLO829BL were prepared and hybridized with the biotin-labelled custom targeted probes following the manufacturer's protocol (Twist Biosciences catalog IDs: 100253, 100255, 100527, 100400). Using streptavidin beads the hybridized DNA was pulled from the DNA pool, and amplified by PCR. Enriched targeted libraries were sequenced on the Illumina NextSeq platform. NextSeq-Capture validation sequencing data were mapped with BWA mem (v0.7.5)[340] and SV calling was performed with Manta[262], independently for COLO829 and COLO829BL. SV calls for COLO829 and COLO829BL were merged using SURVIVOR (v1.0.6, overlap distance of 50bp[343] and only calls with no evidence in COLO829BL were considered as somatic and validated.

PCR: We selected 88 high-confidence SV candidates for PCR validation based on an initial screening of the somatic SV truth set with IGV and added 296 randomly select-

ed additional SV candidates for a total of 384 assays. We automatically designed primers for these SV breakpoints using Primer3 (v1.1.4)[349]. PCR assays were performed on COLO829 and COLO829BL genomic DNA. Libraries were prepared for PCR results and sequenced on both the MiSeq and ONT-MinION platforms. MiSeq-PCR validation sequencing data were mapped with BWA mem (v0.7.5)[340] and SV calling was performed with Manta (v0.29.5)[350], independently for COLO829 and COLO829BL. ONT PCR validation sequencing data were mapped with minimap2 (v2.15[344], and SV calling was performed with NanoSV (v1.2.2)[204] independently for COLO829 and COLO829BL. Moreover, 70 additional SV calls that were shown as somatic in the Capture validation set were also subjected to PCR and products were sequenced on the MinION through the same protocol described above.
SV calls for COLO829 and COLO829BL from the Miseq-PCR and the two Nanopore-PCR sets were merged using SURVIVOR (v1.0.6, overlap distance of 50bp)[343]. Only SV calls with no evidence in any of the COLO829BL sets were considered somatic and validated.

FISH: For FISH validation, we selected probes that bind to 6 genomic regions, including Chromosome Enumeration Probes (CEP) for the centromeric region of chromosome 13, 16 and 18 (CEP13, CEP16, CEP18), labeled with SpectrumOrange (Abbott Vysis, Downers Grove, IL) and centromeric region of chromosome 9 (CEP9), labeled with SpectrumAqua (Leica Biosystems, Amsterdam). Furthermore, locus specific break-apart probes for chromosome 2p23 fusion (SpectrumOrange/SpectrumGreen, Vysis ALK Break Apart, Abbott Vysis, Downers Grove, IL) and 9p24 fusion (SpectrumOrange/SpectrumGreen Leica Biosystems, Amsterdam) were used. COLO829 cells were dissociated using trypsin, counted, washed and diluted to contain a total of 50,000 cells in 100 µl. Monolayer cell suspensions were concentrated on a microscope slide using cytospin. Then, FISH was performed according to diagnostic standards. Slides were visualized on a Leica DM5500 fluorescence microscope and for each probe, 100 cells/slide were recorded.

SV SELECTION PIPELINE

Merged somatic SV calls were overlapped with the validation outcomes with SURVIVOR (v. 1.0.6)[343] using an overlap distance of 50bp (PCR, CAPTURE) and 1kbp (BNG). Only somatic SV calls with support from multiple datasets and calls with support from a single dataset which were validated were selected. SVs involving unstable microsatellites were not considered as part of our analyses. All calls were manually curated by using the SV-plaudit cloud based framework [351] that uses Samplot to generate images from SV coordinates and BAM files. We generated such images for the somatic SV calls for each dataset (ILL, ONT, PB and 10x) and for the validations (PCR-ONT, PCR-MISEQ and

CAPTURE). We evaluated each of these image datasets independently and classified each somatic SV call as "somatic", "germline" or "false positive". We also used the Integrated Genome Viewer (IGV, v2.4.0)[352] to verify some SVs. We performed the same analysis on 176 randomly selected SV calls belonging to a single dataset and which were not validated. Finally, we gathered the somatic SV calls and generated the final somatic VCF file.

COMPARISON TO EXTERNAL SOURCES

CNA calls from Arora *et al.*[214] were downloaded (HiSeq dataset, https://www.nygenome.org/bioinformatics/3-cancer-cell-lines-on-2-sequencers/) and lifted to GRCh37 genomic coordinates with liftOver (UCSC). CNA calls from the four different single cell clusters were obtained from Velazquez-Villareal *et al.*[335]. These datasets were then merged using 1MB bins to calculate Pearson's correlation between datasets and for plotting.

The two somatic SV sets from Arora *et al.*[214] (HiSeq and NovaSeq sets, https://www.nygenome.org/bioinformatics/3-cancer-cell-lines-on-2-sequencers/) were downloaded. Since these are BEDPE files based on GRCh38 genomic coordinates, we converted our somatic SV truth set to BEDPE format and lifted it to those coordinates using the liftOver tool from UCSC. We then intersected those SV sets with our truth set using Bedtools (v2.25.0)[346] and differentiated between SVs with overlap on both sides, overlap only on one side and not overlapping. We lifted all SVs with no overlap or one-sided overlap and manually evaluated them in our data using IGV (v2.4.0)[352].

TUMOR PURITY AND SEQUENCING DEPTH ANALYSIS

For tumor purity simulations in each of the ILL, ONT and PB datasets, COLO829 and COLO829BL BAM files were randomly subsampled and mixed in different ratios, dependent on the sequencing depth to achieve *in silico* tumor purities of 10, 20, 25, 50 and 75 with Sambamba (v0.6.5)[345]. The same somatic SV calling pipeline used for the different datasets was applied to each of the tumor purity subsets. The resulting somatic SV file of each tumor purity subset was overlapped using a window of 100bp with the truth set VCF to determine the number of true and false positives and true negatives. This experiment was performed in triplicate for each tumor purity and each technology with the original COLO829 BAM file as positive control (100% tumor purity) and the original COLO829BL BAM file as negative control (0% tumor purity).

For sequencing depth simulations using the ILL dataset, one of the triplicates from each tumor purity simulation was selected together with the COLO829 and COLO829BL files. Each of these BAM files was subsampled to depths of 1x, 5x, 10x, 30x and 50x (plus the original 98x) with Sambamba (v0.6.5)[345]. Somatic SV calling was performed independently for each of the subsets and the resulting somatic SV VCF file was overlapped with the truth set to determine the number of true and false positives and false negatives.

DATA AVAILABILITY

Genomic data is available on EGA project PRJEB27698; Raw, somatic and truth set VCF files, and CNA files are available in Zenodo DOI: 10.5281/zenodo.3988185;

CODE AVAILABILITY

All code used in the preparation of the somatic SV truth set is available at: https://github.com/UMCUGenetics/COLO829_somaticSV. The code used for simulations of tumor purity and sequencing depth is available at: https://github.com/UMCUGenetics/tumps *Figure panels 2-A, 3-C and 3-D were created using Biorender.com.*

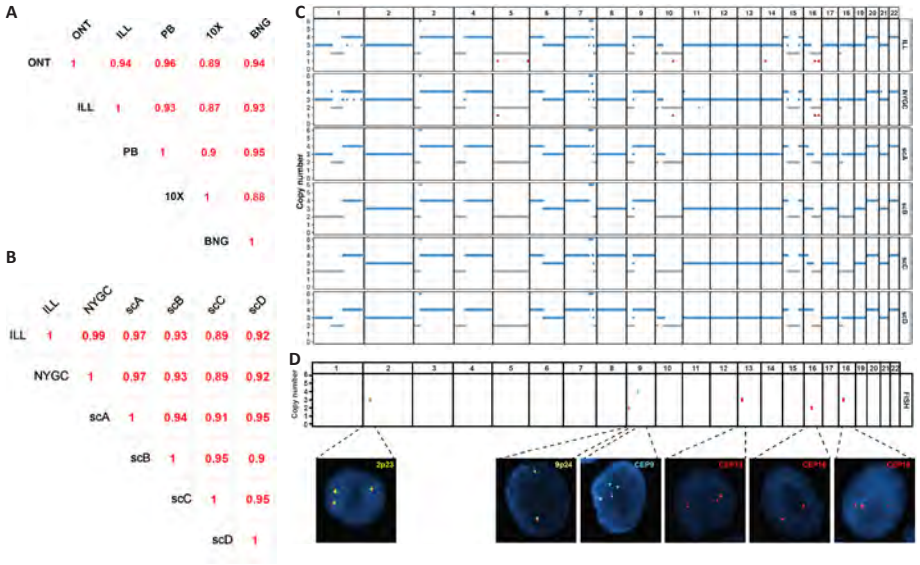
CONFLICT OF INTEREST STATEMENT

AWCP is an employee of Bionano Genomics. AMW is an employee and shareholder of Pacific Biosciences.

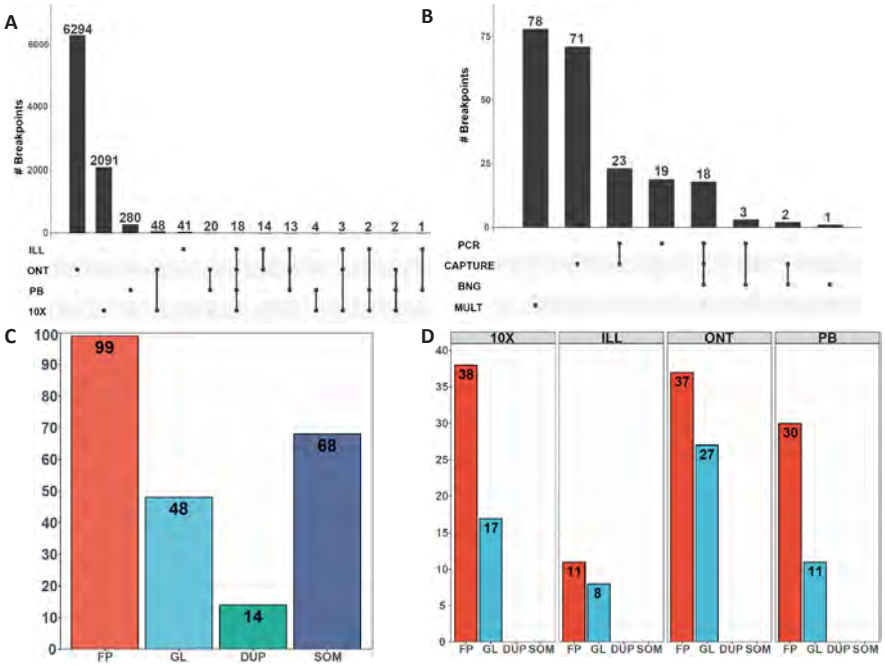
LIST OF SUPPLEMENTARY DATA

Figure S1.	Copy number correlation within our datasets and external datasets
Figure S2.	Generation of a somatic SV truth set
Figure S3.	Characterization of the somatic SV truth set
*Table S1.	Dataset metrics
*Table S2.	Manual curation results
*Table S3.	Truth set annotated
*Table S4.	CNAs annotated
*Table S5.	Arora unique SVs annotated
*Table S6.	Data accession details
*Movie S1	Reconstruction of the breakage-fusion-bridge event in chromosome 3

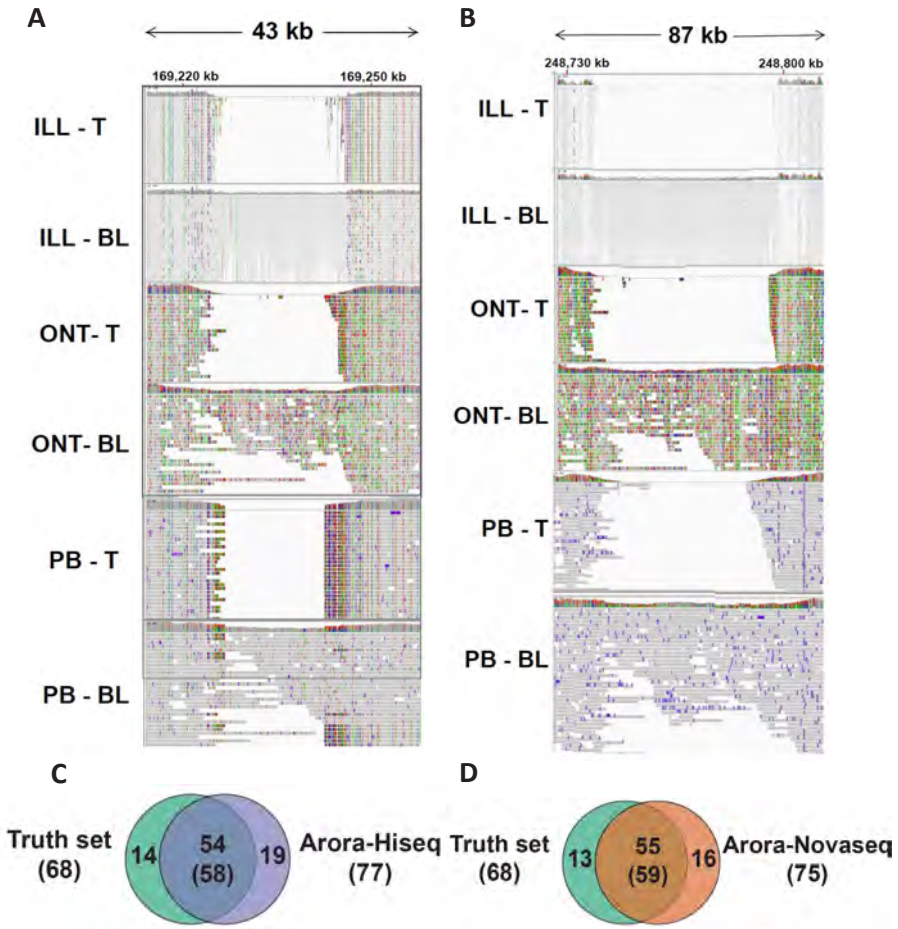
**Table S1-6 and Movie S1 are available online at: https://tinyurl.com/Ch4Suppl or scanning the QR code below*



Supplementary Figure 1 (related to figure 1): Copy number correlation within our datasets and external datasets: Correlation index of CNA calls for (**A**) each of the pairwise comparisons of the datasets generated in our study and (**B**) the comparison of our ILL dataset and the external sets from bulk sequencing in NYGC [214] and the 4 clusters differentiated by single cell sequencing (scA-D) [214,335]. (**C**) Copy number profile of the ILL and the external sets. (**D**) Copy number status of 6 distinct genomic locations as determined by FISH



Supplementary Figure 2 (related to figure 2): Generation of a somatic SV truth set. (**A**) Intersection of the total of 8,831 candidate SV calls merged from all platforms used and presence per in the raw call set per technology. (**B**) Number of validated somatic SV calls per validation approach including multi technology support (MULT). Manual curation statistics for (**C**) validated or multi-dataset SV calls and (**D**) non-validated and single-dataset SV calls. FP = false positive, GL = evidence in germline, DUP = duplication of an already called SV, SOM = real somatic variant.



Supplementary Figure 3 (related to figure 3): Characterization of the somatic SV truth set. (**A**, **B**) IGV screenshots of mapped reads from the ILL, ONT and PB datasets for COLO829 (T) and COLO829BL (BL) of two CNAs on chromosome 1 without associated somatic SVs in the truth set. Overlap of somatic SV calls between our truth set and the two somatic SV sets reported by [214], the Hiseq set (**C**) and the Novaseq set (**D**). One-sided overlaps (i.e. when only one breakpoint of the SV overlaps) are included on the overlap. Numbers in parenthesis indicate the overlap from the Arora set point of view.

Supplementary Movie: Reconstruction of the breakage-fusion-bridge event in chromosome 3. Animated reconstruction of a breakage-fusion-bridge event consistent with the breakpoints and copy number profile of chromosome 3 in COLO829. This circos plot shows the evolution in time and over various cell divisions of the chromosome involving 4 cycles of breakage-fusion-bridge followed by a genome doubling event. The innermost track shows minor allele ploidy (orange indicates loss, blue amplification). The next track shows the copy number profile (purple indicates loss, green amplification). The line track shows the reconstructed chromosome. Breakpoints are represented by triangles and connecting arcs, telomeric ends of the chromosome by squares, and unrepaired double-stranded breaks by circles. DNA gained by replication and new breakpoints formed through DNA repair are indicated in blue, with lost DNA in orange. The outer track shows chromosome number and coordinate. A non-linear chromosomal coordinate scale is used with distances between breakpoints shown in black overlaying the copy number track. A cell cycle clock is shown in the upper left corner indicating at what point in the cell cycle each rearrangement occurs. The final stabilising repair to the centromere of another chromosome is omitted for clarity. Available online at https://doi.org/10.1101/2020.10.15.340497.

ACKNOWLEDGEMENTS

We thank Pacific Biosciences and BioNano for their kind support generating and analysing data. JEV-I is supported by the Gieskes Strijbis Foundation (1816199). This work was performed as part of the EU-funded Horizon2020 EUCANcan project (funding to EC) and the Netherlands X-omics Initiative funded by NWO, project 184.034.019.

AUTHOR CONTRIBUTIONS

Conceptualization: JEV-I, BY, RJAF, WPK, EC
Provided material or generated data: NJMB, EdB, MN, IR
Data analysis: JEV-I, DC, JE, JK, SvL, TM, PP, AWCP, MvR, AMW
Validation: NJMB, IR, MGMR, MvR
Writing manuscript: JEV-I, WPK, EC
Project supervision: WPK, EC

5

OPTIMIZING NANOPORE SEQUENCING-BASED DETECTION OF STRUCTURAL VARIANTS ENABLES INDIVIDUALIZED CIRCULATING TUMOR DNA-BASED DISEASE MONITORING IN CANCER PATIENTS

Optimizing Nanopore sequencing-based detection of structural variants enables individualized circulating tumor DNA-based disease monitoring in cancer patients

Jose Espejo Valle-Inclan[1,2*], Christina Stangl[1,2,3*], Anouk C. de Jong[4*], Lisanne F. van Dessel[4], Markus J. van Roosmalen[1,5], Jean C.A. Helmijr[4], Ivo Renkens[1], Roel Janssen[1,2], Sam de Blank[1], Chris J. de Witte[1,2], John W.M. Martens[4], Maurice P.H.M. Jansen[4], Martijn P. Lolkema[4,$], Wigard P. Kloosterman[1,6,7,$]

[*]These authors contributed equally to this work
[$]corresponding authors

5

[1] Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht and Utrecht University, Utrecht, The Netherlands
[2] Oncode Institute
[3] Division of Molecular Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[4] Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[5] Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands
[6] Cyclomics, Utrecht, The Netherlands
[7] Frame Cancer Therapeutics, Amsterdam, The Netherlands

Submitted and available at: medRxiv 19011932 (2019);
https://doi.org/10.1101/19011932

ABSTRACT

Somatic genomic structural variations (SVs) are promising personalized biomarkers to quantify circulating tumor DNA (ctDNA) in liquid biopsies as they represent unique tumor derived molecules. However, in most solid malignancies these SVs are variable and can be located anywhere in the genome thus the complexity of the identification of personalized SVs hinders routine use in the clinic. Here, we developed a novel approach for rapid discovery of a set of patient-specific somatic SVs. We combine low coverage cancer genome sketching using Oxford Nanopore sequencing with a machine learning approach to detect a set of somatic SVs. We analyzed tumor samples of high-grade ovarian and prostate cancer patients, successfully identified candidate SVs and validated on average ten somatic SVs per patient with breakpoint-spanning PCR mini-amplicons. These SVs could be quantified in ctDNA samples of patients with metastatic prostate cancer using a digital PCR assay. The SV quantification in these longitudinal samples suggest that indeed SV dynamics correlate with and may improve other response biomarkers such as PSA. Our work enables rapid and cost-effective identification of a set of patient-specific SVs that can be used to study ctDNA dynamics.

BACKGROUND

The detection of cancer recurrence as well as accurate and fast monitoring of response to treatment currently lacks sensitivity for detection of changes over time[353,354]. Liquid biopsies, which can be used to detect circulating tumor DNA (ctDNA) from body fluids, such as blood, in a minimally invasive manner, are a promising approach to improve monitoring of tumor burden over time[52,355]. Circulating tumor DNA, which originates from apoptotic and necrotic tumor cells, has been shown to have a positive linear correlation with tumor burden[356]. In multiple cases, ctDNA analysis identified cancer recurrence months before clinical symptoms presented[54,69,165].

As ctDNA is only a fraction of the total circulating cell free DNA (cfDNA), it should be distinguished from cfDNA from normal cells by identification of ctDNA-specific genetic alterations. Genomic structural variations (SVs) represent tumor- and ctDNA-specific biomarkers to detect and quantify ctDNA with high sensitivity in liquid biopsies[69,165,166,357]. Most solid cancers contain dozens to hundreds of somatic SVs[29,150]. Besides some recurrent driver SV events that functionally impact tumorigenesis, the vast majority of these somatic SVs are patient- and tumor specific passenger events[146], which may nevertheless be good biomarkers for tumor load tracing. SVs form a unique breakpoint junction between two joined DNA strands and can be validated by straightforward junction-spanning (quantitative) PCR assays, which facilitates its applicability[165].

Somatic SVs are commonly detected with short-read, paired-end next generation sequencing (NGS). However, as SVs can be very large, short reads are less suited for SV detection[199,200,358]. Recently, long-read sequencing techniques from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) have emerged and their increased power for germline and somatic SV detection has been extensively demonstrated [199,200,203,204,208]. Moreover, ONT enables a short turnaround time and real-time data analysis[188].

To enable rapid and cost-efficient identification of a set of patient-specific somatic SVs for ctDNA monitoring, we developed a pipeline that leverages the long-read and fast sequencing capabilities of nanopore sequencing in combination with a computational method that enables accurate selection of a subset of somatic SVs from low coverage nanopore sequencing data. The method detects a subset of genomic SVs and can be applied to tumor tissue obtained from (needle) biopsy or resection. The computational approach combines SV calling with random forest classification and germline SV filtering against a blacklist to enrich for somatic SVs without the need of matching germline sequencing data, which reduces the cost and time of the assay. We were able to design

SV-specific PCR-assays for ctDNA tracking within three days after obtaining a tumor biopsy. We validated the pipeline in multiple ovarian and prostate cancer samples. In addition, we demonstrate the clinical applicability of our pipeline by retrospectively tracking the identified somatic SVs in longitudinal cfDNA samples of patients with metastatic prostate cancer, by using digital PCR.

RESULTS

DETECTION OF SOMATIC STRUCTURAL VARIATIONS FROM LOW COVERAGE NANOPORE SEQUENCING OF TUMOR BIOPSIES

The first step of our analysis involves low coverage nanopore sequencing of genomic tumor-derived DNA (**Figure 1A**). A single nanopore run on the MinION or GridION platforms typically generates between 5-15 Gbs of data[183], corresponding to 1.5-5x coverage of the human genome. Next, the low coverage sequencing data are mapped to the reference genome followed by the detection of SV breakpoint junctions from split read mappings (**Figure 1B**)[204]. Subsequently, a classification and filtering pipeline is applied to enrich for somatic SV breakpoints irrespective of corresponding germline data (**Figure 1B**). Finally, PCR assays with mini-amplicons are designed to validate the 20 most likely somatic SVs. SVs are confirmed as either somatic or germline by breakpoint PCR on tumor and corresponding lymphocyte DNA (**Figure 1C**). Successful breakpoint PCR assays for somatic SVs can then be utilized as biomarkers for ctDNA-based monitoring of treatment response and disease recurrence (**Figure 1D**).

ESTABLISHMENT OF A SOMATIC SV REFERENCE SET

To verify the ability of our pipeline to detect somatic SVs, we used genomic data from the melanoma cell line COLO829[333] and the ovarian cancer organoid line HGS-3[278]. We utilized short-read WGS data from both lines (90x and 30x coverage for COLO829 and HGS-3, respectively) and matching reference samples (30x coverage in both cases) to establish two reference sets of somatic SVs (Methods). By using a state-of-the-art somatic SV detection pipeline[330,337,359,360], we detected 92 and 295 somatic SVs in COLO829 and HGS-3, respectively. Additionally, we generated long-read nanopore sequencing data for COLO829 and HGS-3, reaching high coverages of 59x (COLO829) and 56x (HGS-3) (**Suppl. Figure 1 and Suppl. Table 1**). To simulate low coverage long-read sequencing of tumor genomes, we randomly subsampled the nanopore sequencing reads to coverages of 4x, 3x and 2x. The subsampling was performed 20 times independently for each case,

to mitigate the effect of chance on the subsampling and subsequent analysis.

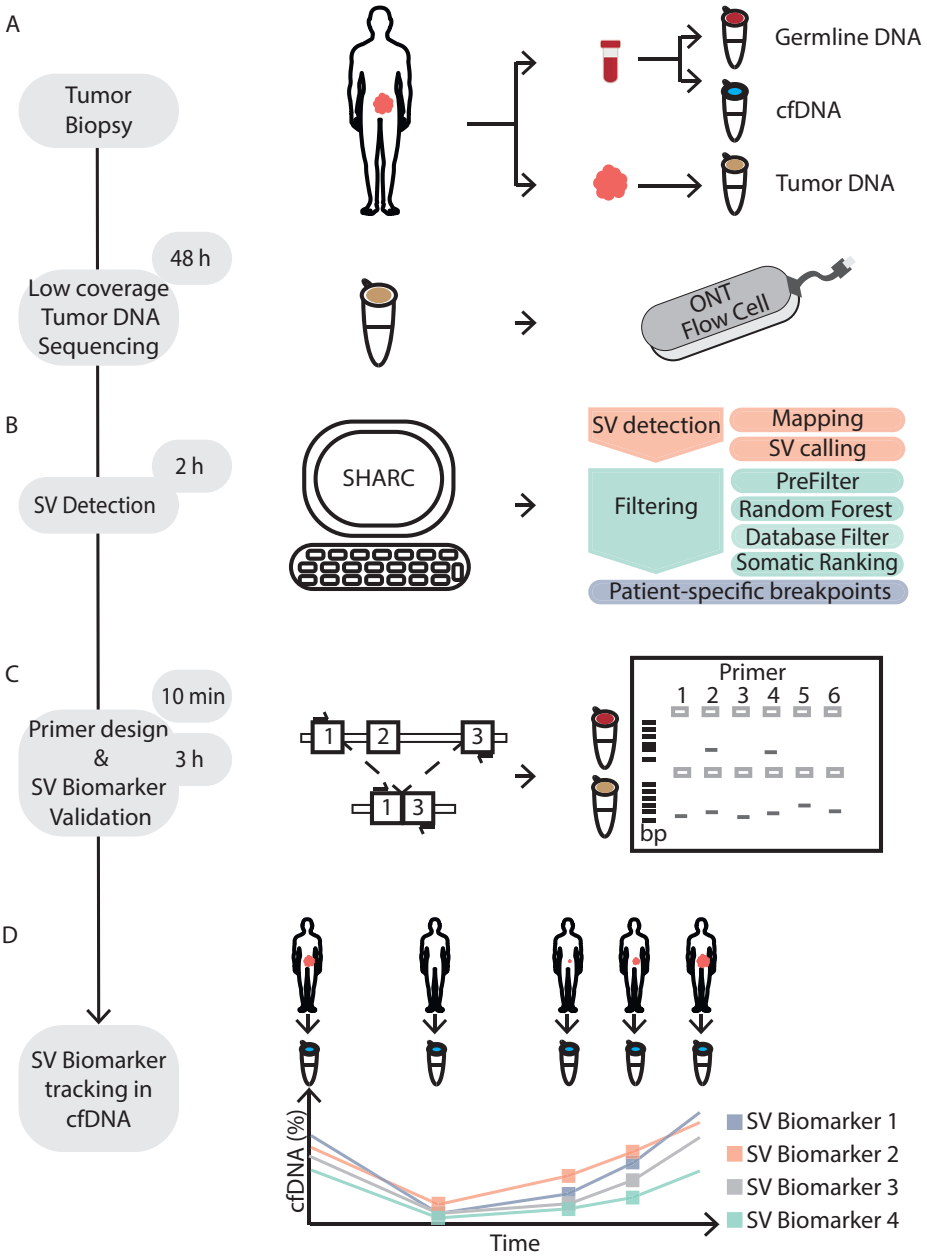


Figure 1: Schematic overview of SHARC. (A) (Needle) biopsy or resection from a tumor as well as blood are obtained from a patient at initial diagnosis. Germline DNA (red) and cfDNA (blue) isolated from blood and tumor DNA (brown) from tumor material. Tumor DNA is sequenced on one ONT flow cell. **(B)** Tumor-specific SV detection and filtering is performed with the bioinformatic SHARC pipeline. **(C)** SV-specific breakpoint spanning primers are designed. Breakpoint PCR with SV-specific primers is performed on

germline and tumor DNA to confirm somatic SVs. (**D**) Somatic SVs are used as biomarkers and traced within cfDNA from a patient to monitor disease dynamics in a longitudinal manner.

Next, we tested our ability to detect SVs from high and low coverage nanopore sequencing data. We used NanoSV, a previously validated nanopore SV caller[203,204], to call SVs from the nanopore sequencing data. To maximize sensitivity, we performed SV calling using lenient settings on high and low coverage COLO829 and HGS-3 Nanopore datasets. (**Suppl. Table 2**). Based on the overlap with the somatic short-read reference set, raw SV calls were classified as somatic (true-positives) or non-somatic (false-positives). As expected, the vast majority of the raw SV calls in all the different coverage datasets were non-somatic, on average 99.84% (range 99.81-99.9%, COLO829) and 99.55% (range 99.4-99.74%, HGS-3) (**Figure 2A**).

In the high coverage Nanopore datasets, we validated 84 (91% of the short-read reference set) and 219 (74% of the short-read reference set) true-positive somatic SVs for COLO829 and HGS-3, respectively, representing a small fraction of the total number of raw SV calls (**Figure 2A and Suppl. Figure 2A**). Similarly, we identified an average of 23 (25% of the short-read reference set) and 53 (18% of the short-read reference set) somatic SV breakpoints in each of the low coverage Nanopore sequencing datasets for COLO829 and HGS-3, respectively. (**Figure 2A**). Furthermore, we compared the performance of the SV callers NanoSV, Sniffles[342] and NanoVar[205]. Thus, we show that based on lenient SV calling of high- and low-coverage Nanopore sequencing data with NanoSV, somatic SVs can be identified.

ENRICHMENT FOR SOMATIC SV CALLS FROM NANOPORE SEQUENCING DATA

Since the somatic SVs identified among the SV call sets of the Nanopore data represent only a small fraction of the total raw SV calls, we implemented a panel of cumulative filtering steps to enrich for somatic SVs. First, we selected only "PASS" SV calls (based on default NanoSV filter flags[204], Methods). Secondly, we excluded calls involving chromosome Y or the mitochondrial genome. Finally, we removed all insertions, since the exact inserted sequence cannot be accurately defined from low coverage nanopore sequencing data, thus hampering the final PCR assay development at a later step. As a result of these filtering steps, 72.6% (COLO829) and 76.2% (HGS-3) false-positive calls were removed in the high coverage sets (**Figure 2B and Suppl. Table 2**). For the low coverage sets, the filtering removed on average 50.9% (COLO829) and 49.9% (HGS-3) of false-positive calls (**Figure 2B and Suppl. Table 2**). In contrast, the vast majority of true-positive somatic SV calls were maintained following SV filtering (on average 76.9% in COLO829 and 93.9% in HGS-3, **Figure 2B**).

To further reduce the number of false-positive SV calls, we employed a random forest (RF) machine learning approach (Methods), similarly as previously described for SV calling of nanopore data[204]. We applied the RF classifier to the filtered high and low coverage subsets of COLO829 and HGS-3. For the high coverage sets, the RF labelled 84% (COLO829) and 81.3% (HGS-3) of false-positive SV calls as false (**Figure 2C**). For the low coverage sets, on average 70.6% (COLO829) and 68% (HGS-3) of false-positive SV calls were labelled as false (**Figure 2C**). In addition, in the high coverage sets, 81.25% (COLO829) and 97.88% (HGS-3) of true-positive somatic SV calls were labelled as true. Similar percentages of true-positive SV calls were labelled as true in the low coverage sets, on average 73.7% (COLO829) and 98.6% (HGS-3) (**Figure 2C**). These results show that the RF classifier filters out the majority of non-somatic breakpoints, while maintaining true-positive somatic SV calls. However, germline SV calls are also maintained at this step, requiring further filtering to enrich for somatic SVs (**Suppl. Figure 2B**).

To reduce the number of germline SVs, we implemented a blacklist filtering step. Therefore, the remaining SV calls were overlapped with two databases (DBFilter) as panel-of-normal (PON) filtering: (i) SharcDB, containing SV calls from nanopore sequencing of 14 different samples , and (ii) RefDB, containing germline SV calls from 59 control samples previously sequenced using Illumina WGS in our group (Methods). Following this filtering step, 100% of true-positive somatic SV calls from both the COLO829 and HGS-3 high and low coverage sets were retained (**Figure 2D**). In contrast, 88.6% (COLO829, high coverage), 76.2% (HGS-3, high coverage) and on average 89.9% (COLO829, low coverage) and 84.5% (HGS-3, low coverage) of remaining false-positive SV calls were filtered out (**Figure 2D**). Due to this filtering, the fraction of true-positive somatic breakpoints among the remaining SV calls increased to 6.6%-18.7%, for the low and high-coverage Nanopore datasets of COLO829 and HGS-3 (**Figure 2E and Suppl. Figure 2A**).

To further enrich for somatic SVs, we implemented a ranking method, based on the observation that large SVs are more likely to be somatic than germline SVs (**Suppl. Figure 4**). This increased the percentage of true-positive somatic SVs to 85% (COLO829) and 65% (HGS-3) in the high coverage sets, and to on average 43% (COLO829) and 64.1% (HGS-3) in the low coverage sets (**Figure 2E**).

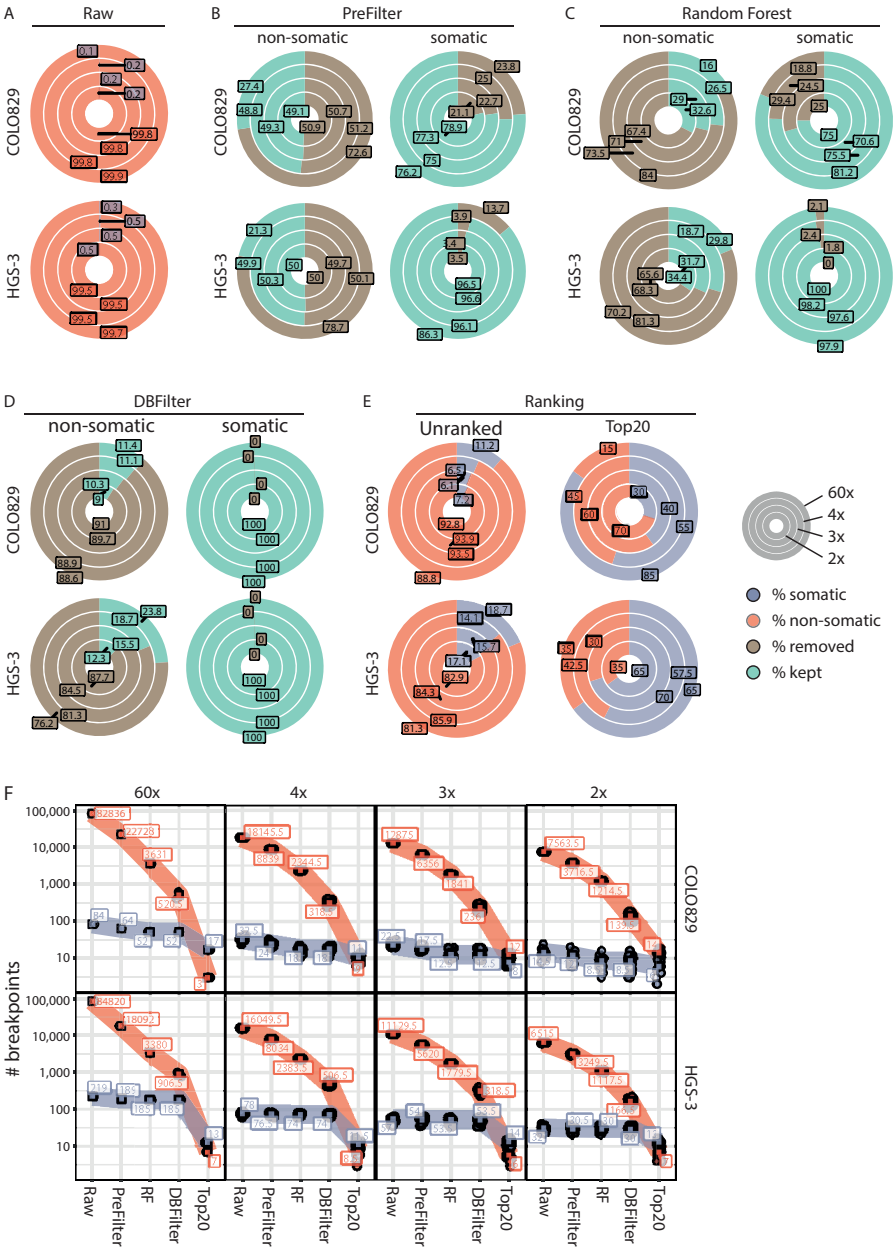


Figure 2: Detection of somatic SVs with the SHARC pipeline based on high and low coverage nanopore data. High coverage nanopore sequencing data from COLO829 (melanoma cell line) and HGS-3 (ovarian cancer organoid) were subsampled to low coverages. Outer circles represent the high coverage sets (59x for COLO829 and 56x for HGS-3) and inner circles *(continues on next page)*
(Figure 2 legend continued) represent low coverage subsets (4x 3x, 2x). The following filtering steps were applied in a cumulative manner in the order displayed. **(A)** Median percentage of non-somatic (red) and somatic (blue) breakpoints in the raw NanoSV calls for COLO829 (top) and HGS-3 (bottom). **(B)** Median

percentage of non-somatic (left) and somatic (right) SV calls kept (green) or removed (brown) in the pre-filtering step for COLO829 and HGS-3. (**C**) Median percentage of non-somatic (left) and somatic (right) SV calls kept (green) or removed (brown) by the Random Forest SV classifier for COLO829 and HGS-3. (**D**) Median percentage of non-somatic (left) and somatic (right) SV calls kept (green) or removed (brown) by the database filtering for COLO829 and HGS-3. (**E**) Median percentage of non-somatic (red) and somatic SV (blue) calls in the complete SHARC output (left) and top 20 largest SVs (right) for COLO829 and HGS-3. (**F**) Total number of non-somatic (red) and somatic (blue) SV calls at each step of the pipeline for both COLO829 and HGS-3. In low coverage subsets, all data points are shown and the square box represents the median value. RF: Random forest; DBFilter: Database filter.

Altogether, our SV filtering pipeline strongly enriches for true-positive somatic breakpoints and filters out the majority of false-positives and germline SVs. We demonstrate a total enrichment of true-positive somatic SV calls from 0.1% in the raw calls to 85% in the final Top20 ranked calls (17/20, COLO829, high coverage), 0.26% to 65% (13/20, HGS-3, high coverage), on average 0.18% to 41.7% (8.3/20, COLO829, low coverage sets) and on average 0.49% to 64.2% (12.8/20, HGS-3, low coverage sets) (**Figure 2F**). Of note, despite low coverage sequencing, each of the somatic SV calls identifies breakpoints at nucleotide resolution, providing immediate access to breakpoint PCR testing.

VALIDATION IN TUMOR TISSUE FROM PATIENTS WITH OVARIAN AND PROSTATE CANCER

Next, we tested the pipeline on four high-grade serous ovarian cancer (Ova1-4) and six prostate cancer (Pros1-6) samples. We sequenced tumor DNA on one nanopore flow cell per sample. The ovarian cancer samples and three prostate cancer samples (Pros1-3) were sequenced on commercial ONT flow cells. For the ovarian cancer samples, we started library preparation with minimally 1 µg of DNA. For the prostate cancer samples limited material was available, and we started library preparation with 250 ng of DNA. For one sample (Pros3), not enough sequencing data was produced to confidently detect somatic SVs and this sample was therefore excluded from all subsequent analyses (**Suppl. Table 1**). Three additional prostate cancer samples (Pros4-6) were sequenced on ONT research prototype flow cells with higher sequencing sensitivity, thus requiring less DNA input material. In these cases, library preparation was started with an average of 108 ng (80-128 ng) of DNA and an average of 10 ng of library was loaded for sequencing (**Suppl. Table 1**). We obtained an average sequence coverage of 2.3x (range: 1.8 - 4.0) (**Figure 3A and Suppl. Table 1**) and average read lengths of 7.8 Kbp (range: 4.2-16.3 Kbp) (**Figure 3B and Suppl. Table 1**). The sequencing throughput was not affected by the lower DNA input when using the high-sensitivity prototype flow cells. (**Suppl. Table 1**).
Following the lenient SV calling, pre-filtering, RF classification, the database filtering and ranking steps, an average 2.8% (range of 1.0%-4.4%) of SVs per sample were retained (**Figure 3C**). We performed breakpoint PCR assays on lymphocyte and tumor DNA for the top 20 ranked SVs and validated an average of 10 (50%, range 25-80%) somatic

SVs per sample (**Figure 3D**). Therefore, despite not having enough sequencing depth to provide a complete genome construction, we were able to identify several somatic SV biomarkers in each of the tumor samples.

We investigated the recall of validated somatic SVs at different timepoints during the sequencing run. We found that, on average, 81.6% (range 50-100%) of validated somatic SVs were already detected within the first 24 hours of sequencing (**Suppl. Figure 6**). This offers the opportunity to reduce the sequencing time, accelerating tumor biomarker discovery with one day.

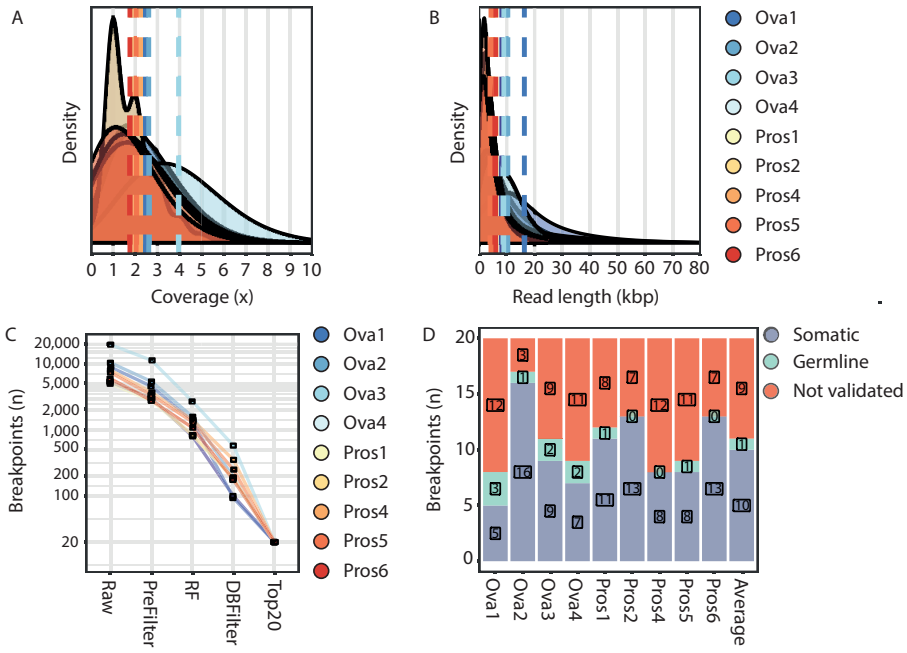


Figure 3: SHARC identifies and validates tumor-specific SV biomarkers from low-pass nanopore tumor sequencing data. Plots showing the distribution of (**A**) coverage and (**B**) read length for the nine tumor samples sequenced on one flow cell each. Dashed lines represent averages for each sample. (**C**) Total number of somatic SVs present at each of the steps throughout the SV calling and filtering pipeline. RF: Random forest; DBFilter: Database filter (**D**). The Top20 ranked breakpoints for each sample were tested by breakpoint PCR using tumor and germline DNA. Graph depicts the number of breakpoints validated as somatic (blue), germline (green) or breakpoints that could not be validated (red).

DETECTION OF SOMATIC SVS IN CFDNA FROM PATIENTS WITH OVARIAN AND PROSTATE CANCER

To show the applicability of the pipeline to detect clinically relevant biomarkers, we next tested if we could detect the validated somatic SVs in cfDNA of patients. Ascites fluid, which is known to contain cfDNA and ctDNA[361] was available for Ova2 at time of disease recurrence. We extracted cfDNA from the ascites and tested the 16 validated somatic SVs out of the Top20 by PCR. 100% of somatic SVs could be detected within the cfDNA from ascites (**Suppl. Figure 7**), and not in the germline or water controls. Next, we tested whether validated SVs could be detected in cfDNA from blood. Therefore, we selected two patient-specific SVs for four prostate cancer patients (Pros1, 4, 5 and 6) based on a high signal to noise ratio observed in qPCR assays for SV breakpoints (**Figure 4A** and Methods).

To enable sensitive and quantitative detection, we designed digital PCR (dPCR) assays for the eight selected SVs (**Figure 4B**). For each SV, we aimed to design a probe for both wild-type alleles (up- and downstream) and for the mutant allele (across the breakpoint junction). For five SVs we could design an assay that quantified both the upstream and downstream wild-type allele. For the three other SVs, primers/probes for only one of the wild-type alleles were designed, as appropriate primer design for the other allele was hindered by repetitive sequences at the target site. As the amount of cfDNA within one liquid biopsy is limited, we used a conditional breakpoint detection approach: (i) if dPCR on pre-amplified cfDNA (input pre-amplification: 0.2-1 ng cfDNA) confirmed the presence of the SV within cfDNA, (ii) then subsequent dPCR on non-preamplified cfDNA (stock cfDNA) (input dPCR: 5 ng cfDNA) was performed. The latter enabled calculation of both the variant allele frequency (VAF) and the number of mutant molecules per milliliter plasma (MM/mL plasma). First, we selected two timepoints per patient, one at baseline and one at progression of disease and confirmed the presence of all eight SVs with dPCR on pre-amplified cfDNA (**Suppl. Figure 8**). Thereafter, dPCR on the stock cfDNA successfully detected all SVs in the four patients, both in baseline and progression samples (**Figure 4C and 4D**). Despite the fact that the VAF in pre-amplified cfDNA correlates to the VAF in stock cfDNA ($r_s = 0.928$), they should be considered two separate outcome measurements (regression coefficient = $0.72 \neq 1$) (**Suppl. Figure 9A**). Moreover, VAF based on the wild-type upstream allele was highly similar to VAF based on the wild-type downstream allele in stock cfDNA ($r_s = 0.996$, regression coefficient = 1.05) (**Suppl. Figure 9B**), suggesting no significant imbalances between the two sides of the breakpoint.

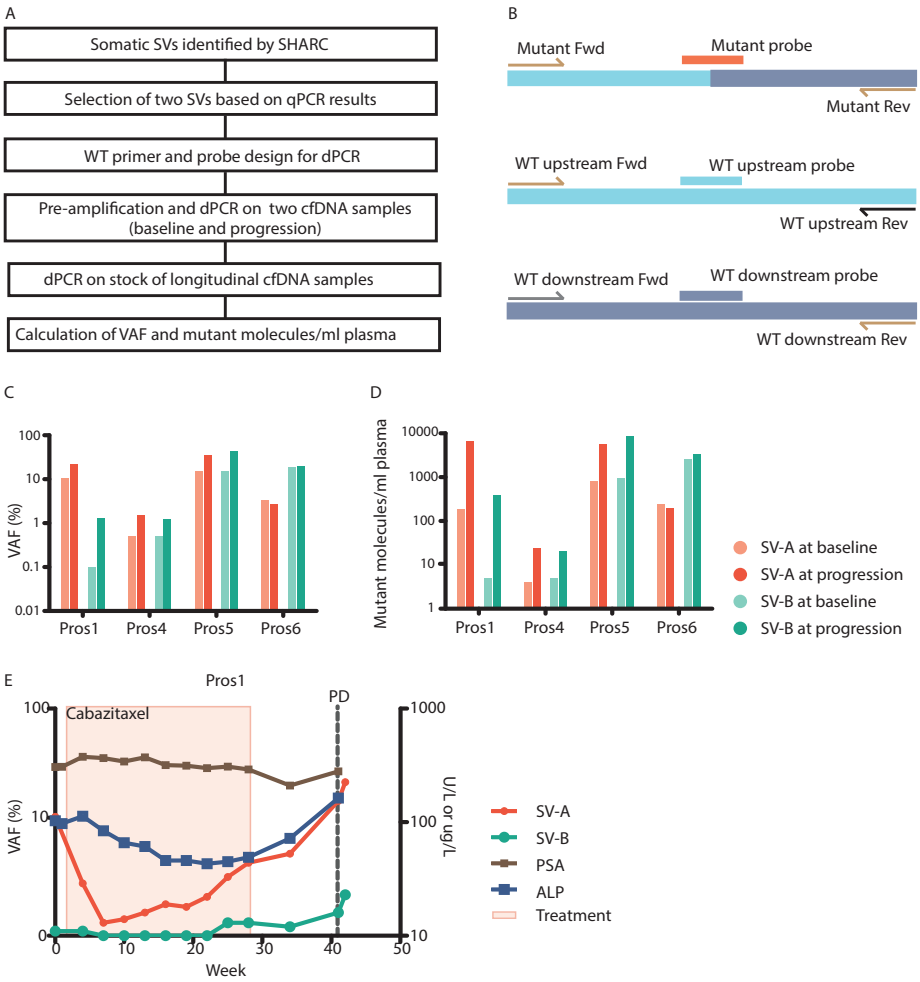


Figure 4: dPCR-based quantification of SVs in blood. **(A)** Schematic overview of quantification of tumor-specific SVs, identified by SHARC, in cfDNA from blood by using qPCR and dPCR. **(B)** Primer and probe design for dPCR. The wild-type upstream and wild-type downstream allele share each one primer with the mutant allele. Three probes with different fluorescents were designed to specifically detect the mutant allele or one of the wild-type alleles. **(C)** Detection of two tumor-specific SVs in cfDNA from blood from four patients with prostate cancer at baseline and at progression of disease with dPCR. Shown are VAF and **(D)** mutant molecules per mL plasma. **(E)** Quantification of SVs in longitudinal cfDNA samples from blood of patient Pros1. Graph depicts VAFs of SVs, treatment, laboratory parameters (prostate specific membrane antigen (PSA), alkaline phosphatase (ALP) and clinical progression of disease (PD).

MONITORING TREATMENT RESPONSE IN PATIENTS WITH PROSTATE CANCER

In addition to the detection of SVs in cfDNA at baseline and progression of disease, we explored the capacity to use SVs to monitor treatment response over time. To enable reliable response monitoring, measurements should be accurate and repeatable. As VAFs are ratios and in principle not influenced by technical variations between timepoints, we chose to report VAFs only. To verify the accuracy of dPCR, we performed two technical replicates for all pre-amplified samples of Pros5 and Pros6 and confirmed a high correlation of VAFs between the replicates (r_s = 0.987, regression coefficient = 0.918) (**Suppl. Figure 9C**). Finally, we quantified the eight SVs of the four prostate cancer patients in the longitudinally collected samples from before, during and after treatment. For Pros1, SV-A shows the potential to improve response evaluation as its dynamics correspond to the expected response to treatment with cabazitaxel and increases towards the end of treatment, resulting in the highest levels at clinical progression of disease (**Figure 4E**). These changes also seem to correlate with other blood biomarkers, including PSA and ALP. In addition, SV-B in Pros1 similarly correlates with response to treatment (**Figure 4E**). Also for Pros5 both SV-A and SV-B show clear changes over time correlating with clinical parameters, and Pros4 and Pros6 have less compelling dynamics of the detected SVs (**Suppl. Figure 10A-C**).

DISCUSSION

Recent studies have utilized somatic SVs for tracking tumor burden from liquid biopsies[69,165,166,357]. Although these studies showed the potential of this methodology, they lacked sufficient turn-around time to provide personalized biomarkers before the initiation of patient treatment. This is due to lengthy short-read WGS approaches for SV detection and an associated substantial number of false-positive somatic SVs, requiring laborious testing to validate SVs. To overcome these limitations, we utilized the real-time and long-read capabilities of nanopore sequencing combined with a machine learning approach to efficiently identify a set of somatic SVs from tumor tissue within three days. The rapid and simple workflow offers great potential for routine monitoring of cancer dynamics. We illustrate the applicability of our method to measure tumor burden by using a series of longitudinally gathered blood samples from metastatic prostate cancer patients.

Obtaining enough tumor material for DNA isolation is often a limiting factor for next-generation sequencing assays. We show that nanopore sequencing and somatic SV detection is possible from limited amounts of DNA that can be extracted from a metastatic tumor needle biopsy, which is an important requisite for clinical viability. DNA in-

put can be decreased even further to as little as 80 ng when using flow cells with increased sensitivity for DNA (research prototype flow cells provided by ONT).
Long-read sequencing is an excellent method for the detection of SVs at nucleotide resolution, even at low sequencing depth, because each long-read that bridges a breakpoint-junction provides direct information on the breakpoint position and sequence [204]. Sequencing of a tumor sample on a single GridION/MinION nanopore flow cell generates insufficient sequencing data to accurately establish a complete genomic profile. However, using the pipeline developed here, we efficiently enriched for patient-specific somatic SV events - irrespective of their functional impact on tumor biology. Despite the very low coverage, the computational method functions independently of corresponding germline sequencing data. These assets make our pipeline a cost-efficient assay for detection of personalized somatic SV biomarkers. Furthermore, on average 50% of the detected SVs are somatic, which minimizes the hands-on effort needed for validation purposes. For all analyzed tumors, we identified at least five somatic SV biomarkers per patient, an amount within the range of biomarkers used to trace ctDNA in previous work[69,166,362]. With expected increases in sequencing throughput from ONT sequencing, the performance of the pipeline will improve significantly. Furthermore, the use of cheap disposable flow cells (Flongle) could reduce assay costs to 1/5 of the current sequencing price of 800€ [363]. The minimal costs of this assay would enable the broader application of such individualized SV monitoring in cancer patients.
We retrospectively traced levels of ctDNA with two SVs per patient for four prostate cancer patients and compared tumor dynamics to clinical biomarkers such as PSA and ALP. The quantitative measurement of SVs in ctDNA suggests that VAFs of SVs correlate with tumor load (Pros1 and Pros5). Moreover, the SVs would have indicated progression of disease earlier than PSA did in some patients (Pros1 and Pros 4). Even though we only tested two SVs per patient, this clearly illustrates the potential clinical utility of quantifying ctDNA with SVs to monitor response to treatment. The assay could be optimized by not only identifying the tumor-specific SVs, but also SVs that represent the dominant disease clone and upcoming, targetable subclones. In addition, larger prospective studies should confirm that indeed measuring SVs improves clinical decision making in patients with metastatic prostate, and other cancer types.

CONCLUSIONS

Clinicians are well aware of the dynamic response of cancer to treatment but lack the tools to monitor these changes in real-time and thus generally respond to alterations too late for true treatment success. We present a method to overcome these limitations and provide a solution to immediate individualized disease monitoring. This approach could

increase sensitivity of disease monitoring to such levels that more intelligent treatment approaches could be envisioned.

MATERIALS AND METHODS

DNA Isolation and nanopore sequencing: COLO829 (ATCC® CRL-1974™) cell line was obtained from the American Type Culture Collection (ATCC) and grown according to standard procedures as recommended by ATCC. DNA was isolated using a phenol chloroform protocol[364]. For some nanopore sequencing runs, DNA was sheared using g-tubes (Covaris). DNA was size selected on the PippinHT (Sage Science). Library preparation was performed using the Lib SQK-LSK109 kit (Oxford Nanopore Technologies) and DNA was then sequenced in 49 separate runs using R9.4 flow cells (Oxford Nanopore Technologies) on the MinION (44), GridION (3) and PromethION (2) instruments (**Suppl. Table 1**).

HGS-3 organoid line was cultured following the ovarian cancer organoid culture protocol[278]. DNA was isolated by using a phenol chloroform protocol[364]. DNA was size selected on the PippinHT (Sage Science). Library preparation was performed using the Lib SQK-LSK109 kit (Oxford Nanopore Technologies) and DNA was then sequenced in 40 separate runs using R9.4 (23) and R9.5 (17) flow cells (Oxford Nanopore Technologies) on the MinION (35) and GridION (5) instruments (**Suppl. Table 1**).

Tumor DNA from patients with ovarian cancer was isolated with the Genomic-tip kit (Qiagen), following the manufacturer's protocol for tissue samples. DNA was prepared for nanopore sequencing with the Lib SQK-LSK109 (Oxford Nanopore Technologies). The library from one tumor sample was loaded on one revD (Ova1) or R9.4 (Ova2-4) flow cell (Oxford Nanopore Technologies). Sequencing was performed on a MinION (Ova2, Ova4) or GridION (Ova1, Ova3) instrument (Oxford Nanopore Technologies) (**Suppl. Table 1**). Lymphocyte DNA for PCR validation assays was isolated from blood with the DNeasy Blood & Tissue Kit (Qiagen).

Tumor and germline DNA from patients with prostate cancer were obtained from a fresh frozen core needle biopsy of a metastatic lesion and blood, respectively. DNA was isolated on an automated setup with the QiaSymphony according to the supplier's protocols (DSP DNA Midi kit for blood and DSP DNA Mini kit for tissue). In the context of the CPCT-02 study, WGS was performed by the Hartwig Medical Foundation, Amsterdam, The Netherlands[365]. Residual tumor DNA (80-250 ng) was used for nanopore sequencing. DNA was prepared for nanopore sequencing with the Lib SQK-LSK109 (Oxford Nanopore Technologies). The library from one tumor sample was loaded on one R9.4 (Pros1), revD (Pros2,3) or high-sensitivity research prototype (Pros4-6) flow cell (Oxford Nanopore Technologies). Sequencing was performed on a GridION instrument

(Oxford Nanopore Technologies) (**Suppl. Table 1**).

Illumina sequencing and analysis (COLO829 and HGS-3): Short read WGS was obtained for matched tumor and normal DNA from the COLO829 cell line[284] and the HGS-3 organoid line[278].
SV calling was performed by using GRIDSS (v. 2.0.1)[318] in joint calling mode (tumor+reference) for COLO829 and HGS-3 separately. Somatic SV calls were filtered as in[284] (https://github.com/hartwigmedical/pipeline/blob/master/scripts/gridss_somatic_filter.R)

Benchmarking somatic SV calling from low coverage nanopore sequencing data: Nanopore data from COLO829 was randomly subsampled to 5x sequencing coverage three times independently with Sambamba[345]. SV calling was performed with NanoSV (v. 1.2.4)[204] with a 2-read support threshold; Sniffles (v. 1.0.12) [342]with parameters "--report_BND --genotype -s 2"; and NanoVar (v. 1.3.8) [205]with default parameters. In all cases 8 threads were used and computational resources were measured with GNU Time. True and false positives were calculated using the short-read somatic SV callset described above.

SV calling and filtering pipeline: The SHARC pipeline is available through https://github.com/UMCUGenetics/SHARC. Mapping is performed in parallel for each FASTQ file by using minimap2 (v. 2.12)[344] with settings "-x map-ont -a --MD". The reference genome used is version GRCh37. Sorting and merging of BAM files was done by using sambamba (v. 0.6.5)[345]. SV calling was performed by using NanoSV (v. 1.1.2)[204]. Default NanoSV settings were used except a minimum read count of 2 (cluster_count=2) and minimum mapping quality of 20 (min_mapq=20). VCFs are filtered by using the command `awk '$7 == "PASS" && $1 !~ /(Y|MT)/ && $5 !~ /(Y|MT):/ && $5 != "<INS>"'` to select PASS calls and remove insertions and SVs involving chromosomes Y or MT. VCFs are then annotated with the distance to the closest single repeat element in the reference genome[366,367], the closest gap element in the reference genome[367,368], and the closest segmental duplication element in the reference genome[367]. These elements were taken from the UCSC genome browser (http://genome.ucsc.edu)[367], using the GRCh37/hg19 genome version.
We trained a random forest (RF) model to filter out false-positive SV calls from nanopore data, similarly as previously described[204]. We expanded the selection of input features for the RF, by including read length, SV calling features, and overlap with repeat features in the reference genome (**Suppl. Table 3**). We trained the classifier on the well-characterized NA12878 Genome in a Bottle (GIAB) sample[183,369,370], for which high-quality germline SV call sets have been obtained by using Illumina[370] , PacBio[369] and Nanopore[183] se-

quencing. The GIAB SV truth set was generated by intersecting these three GIAB SV sets resulting in a set of 1,515 germline SVs. We used ⅔ of the GIAB truth set as a training set and ⅓ as a test set. We established a precision-recall curve from 100 bootstrapping runs (**Suppl. Figure 4**), where the training data were split into 90%-10% train-test subsets. Based on the precision-recall curve, we defined an operating point of 96% precision and 99.5% recall (**Suppl. Figure 4**). The final model was then re-trained on the whole training set and tested on the ⅓ test set. The performance on the test set was 95.1% precision and 99.6% recall, representing an accuracy of 97.2% (**Suppl. Figure 4**). SV candidates are classified as "true" or "false" based on this RF model.

We set up two databases of SV calls: (i) SharcDB: containing raw NanoSV calls from nanopore sequencing data of 14 samples, 11 of which belong to this study (COLO829, HGS-3, Ova1, Ova2, Ova3, Ova4, Pros1, Pros2, Pros4, Pros5 and Pros6; and three more for which we had SV calls from high coverage nanopore data: COLO829BL (lymphoblastoid cell line, 50x sequencing depth), VCAP (prostate cancer cell line[371]) and the Genome in a Bottle SV calls (GIAB, [183], . For tests performed with the samples included in this study, the specific sample was excluded from blacklisting with SharcDB; (ii) RefDB: containing germline calls obtained from WGS short-read data of 59 controls: 19 blood controls from patients with ovarian cancer[278], where germline SVs were called with Manta (v. 1.0.3)[350] with default parameters and 40 healthy individuals (biological parents of individuals with congenital abnormalities)[372] where germline SVs were called with Manta (v. 0.29.5)[350] with default parameters.SV calls from tumor samples are overlapped with those two databases by using VCF-explorer (https://github.com/UMCUGenetics/vcf-explorer). Only samples classified as "true" by the RF model and that do not overlap with any sample in the databases qualify for primer design. Primer design for filtered SV calls is automatized by using Primer3 (v. 1.1.4)[349] with a product size range of 30-230 bp. SVs with a successful primer design are ranked based on SV length and the 20 largest are selected for PCR validation. Insertions are filtered out early in the pipeline since the inserted sequence cannot be accurately inferred from the low coverage nanopore sequencing data. Inter-chromosomal translocations are not present in the top20 ranked SVs because the final ranking is based on SV size and this cannot be determined for inter-chromosomal SVs. However they are available in the final VCF file and primers are designed by default, so they can be manually selected for PCR validation and assay development.

Breakpoint PCR: To validate SVs, breakpoint PCR with AmpliTaqGold (Applied Biosystems) was performed according to the manufacturer's protocol. 10 ng primary tumor DNA (somatic) and 10 ng lymphocyte DNA (germline) per primer-pair were used as input. PCR products were loaded and visualized on a 2% agarose gel.

cfDNA isolation: cfDNA was isolated from ascites fluid of Ova2 by using the QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's protocol. Plasma samples from patients with prostate cancer were obtained longitudinally during treatment in 3x10 ml CellSave preservative tubes (Menarini Silicon Biosystems, Huntingdon Valley, PA, USA) and processed within 96 hours as previously described[373]. Circulating DNA was isolated with the QIAsymphony® DSP Circulating DNA Kit (Qiagen) according to manufacturer's protocol with some minor modifications[374]. All cfDNA samples were quantified by Qubit™ fluorometric quantitation (Invitrogen).

Quantitative PCR: As primer specificity is essential for reliable interpretation of an endpoint assay like dPCR, primers for the detection of structural variants were validated by quantitative PCR (qPCR) on whole genome amplified (WGA) tumor and germline DNA. In brief, qPCR was performed by using the CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad Laboratories) and the final reaction mix consisted of 10 µL SensiFAST™ SYBR® Lo-Rox mix (Bioline), 0.5 µM forward and reverse primers, 10 ng of WGA DNA and Ultrapure DNas/RNAse free H_2O to bring up the reaction volume to 20 µL. The Cycle conditions were as follows: 14 cycles of 10s at 95°C and 30s at from 65-58°C (touchdown), followed by 20-40 cycles of 10s at 95°C and 30s at 60°C. In addition, a melt curve was generated from 56°C to 95°C to assess the generated PCR products. Based on qPCR results, two primer sets for the detection of SVs in each patient were selected for quantification by digital PCR (dPCR). Primer sets were excluded from use with dPCR when one of the following ocurred: >1 PCR product, $Cq_{germline}$- Cq_{tumor} <5 and/or Cq_{tumor} > 20.

DNA sonication and fragment size analysis: To mimic the length of cfDNA and improve DNA molecule partition, WGA DNA of both tumor and germline were sonicated to a peak size of ~150 bp with the S220 Focused-ultrasonicator (Covaris) according to the manufacturer's protocol. The sonication conditions were as follows; 200-250 ng WGA DNA (concentration determined by Qubit™ fluorometric quantitation) in 50 µL Ultrapure DNas/RNAse free H_2O, Peak Incident Power: 175 W, Duty Factor: 10 %, Cycles per Burst: 200, Treatment Time: 280 s, Temperature: 7°C, and Water Level: 12. After sonication DNA fragment sizes were analyzed with the High Sensitivity DNA kit (Agilent Technologies) on the Bioanalyzer (Agilent Technologies) and the sample concentration was re-quantified by Qubit™ fluorometric quantitation (Invitrogen, Life Technologies, Carlsbad, CA, USA).

Design of digital PCR assays for absolute quantification of SVs in cfDNA: To quantify SVs in cfDNA, dPCR was performed. First, the exact position of the breakpoint as determined by nanopore sequencing was validated. We used already available sequenced

Illumina data from the CPCT-02 study (Pros1, Pros4, Pros5 and Pros6), but Sanger sequencing of the particular qPCR product could be used as well. To enable quantification of both mutant and wild-type alleles, additional primers for the detection of wild-type upstream (WT-U) allele and wild-type downstream (WT-D) allele of the breakpoint and fluorescent probes for both mutant and wild-type alleles were developed by using the Primer Express Software v3.0 (ThermoFisher) and the online tool Primer3Plus[349]. All primers and fluorescent probes (**Suppl. Table 4**) were ordered from Eurogentec.

Pre-amplification of cfDNA: To enable sensitive detection of multiple SVs in limited amounts of cfDNA, two SVs per patient were pre-amplified with 0.2-1 ng of cfDNA. Pre-amplified tumor and germline DNA samples were used as respectively positive and negative control. Pre-amplification was performed by using 4 µL of TaqMan™ PreAmp Master Mix (cat.no: 4488593, Life Technologies), 2 µL primer pool (0.25 µM) consisting of SV forward (SV-F) and reverse (SV-R) primers and upstream (WT-U) and downstream (WT-D) wild-type primers, and 2 µL (cf)DNA for a total volume of 8 µL. Pre-amplification cycle conditions were: 10 min at 95°C followed by 14 cycles of 15 s at 95°C and 4 min at 60°C, and finally pause at 4°C. After the pre-amplification reaction, 72 µL of Ultrapure DNase/RNAse free H_2O was added to each sample. Next, pre-amplified cfDNA was diluted 40x per 1 ng input, used for the pre-amplification, to prevent overloading of the dPCR chips.

Absolute quantification of SVs in cfDNA with digital PCR: For the quantification of SVs in (cf)DNA, dPCR was performed with the Naica Crystal PCR system (Stilla Technologies) by using the following optimized reaction mix: 1 µL of diluted pre-amplified (cf)DNA sample, 5.6 µL PerfeCTa Multiplex qPCR ToughMix (Cat.No: 733-2322PQ, Quantabio). 0.25 µM probes (SV^{FAM}, $WT\text{-}U^{HEX}$, $WT\text{-}D^{CY5}$), 0.75 µM of the SV forward (SV-F) and reverse primer (SV-R), 0.25 µM of the WT-U and WT-D primers, 0.1 µM Fluorescein (Cat.No: 0681-100G, VWR) and Ultrapure DNAse/RNAse free H_2O to bring up the total volume to 28 µL. Samples were loaded onto Stilla Sapphire chips (Cat. no. C13000, Stilla Technologies) and dPCR was performed with the same cycle conditions as for the primer validation with qPCR. Median number of analyzable droplets was 21,357, inter quartile range 19,837-22,736. dPCR reactions were optimized with 10 ng sonicated tumor and germline WGA DNA. When an SV could be detected in pre-amplified cfDNA samples, a dPCR of all longitudinal cfDNA samples was performed on 5 ng of stock (no pre-amplification) cfDNA to enable absolute quantification of mutant molecules in plasma.

Statistics: qPCR experiments were analyzed with Bio-Rad CFX Manager version 3.1. dPCR experiments were analyzed with Crystal Miner™ software, version 2.1.6 (Stilla

Technologies). Thresholds for positive fluorescence were determined per primer pair based on positive and negative controls. Variant allele frequency (VAF) was calculated according to the following formula:
number of mutant molecules per µl in chip (as defined by Crystal Miner™ software) / (number of mutant molecules per µl in chip + number of wild-type molecules per µl in chip) * 100%.
Absolute number of mutant molecules per mL plasma was calculated as follows:
number of mutant molecules per µl in chip * 28 µl (input in chip) / (used eluate/total volume of eluate * volume of plasma used for isolation).
To correct for zero values on a log scale, +1 was counted to every value and axes were corrected with -1. Spearman's correlation coefficient was calculated for comparisons of VAF based on upstream wild-type allele vs downstream wild-type allele, two replicates and pre-amplified vs non-pre-amplified cfDNA samples. Corresponding slope was calculated by using linear regression analysis.

DECLARATIONS

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Tumor samples of four patients with high-grade serous ovarian cancer (OC) and six patients with metastatic castration resistant prostate cancer (PC) were used in this study. Patients with OC participated in the HUB-OVI study, in which tumor tissue and blood were obtained for organoid culture (tumor) and whole genome sequencing (WGS) (tumor and blood). Clinical data was extracted from the patient file in collaboration with the Dutch Cancer Registration. Patients with PC participated in both the CPCT-02 study (NCT01855477) and the CIRCUS study (NTR5732), in which tumor tissue from a metastatic lesion for WGS and longitudinal cfDNA samples were obtained. Longitudinal ctDNA quantification was performed for four patients with PC. Informed consent was obtained within all studies. Clinical data for patients with PC were collected in an electronic case report form (ALEA Clinical). All studies were performed according to the guidelines of the European Network of Research Ethics Committees (EUREC) following European, national and local law.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Nanopore sequencing data is available at the European Nucleotide Archive (ENA) and through controlled access at the European Genome-phenome Archive (EGA) as follows:
COLO829 cell line: ENA accession ERX2765498.
HGS-3 organoid line: EGA dataset accession EGAD00001005476
Patient material: EGA study accession EGAS00001003963
Data access requests will be evaluated by the UMCU Department of Genetics Data Access Board (EGAC00001000432)
SHARC SV filtering pipeline is available through https://github.com/UMCUGenetics/SHARC

COMPETING INTERESTS

J.E.V-I, C.S. and W.P.K. have received financial compensation for travel and accommodation expenses to speak at Oxford Nanopore Technologies-organized meetings. The remaining authors declare no competing financial interests.

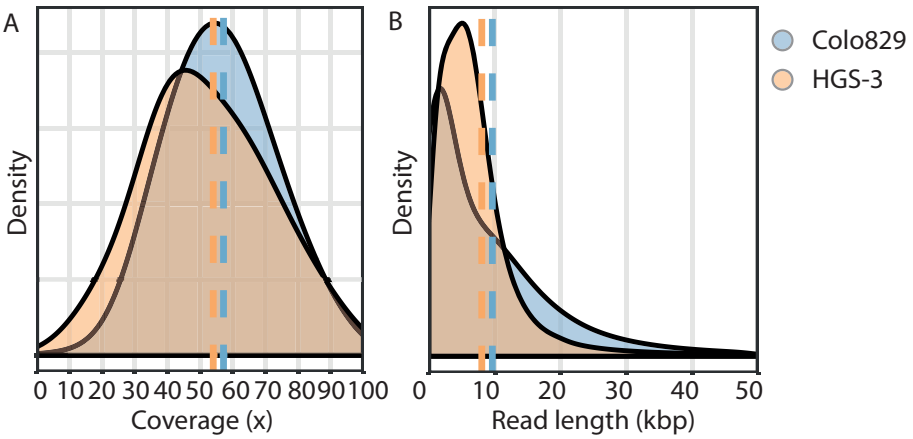
FUNDING

This work has been supported by KWF grants UU 2012-5710 and by funding from the Utrecht University to implement a single-molecule sequencing facility.

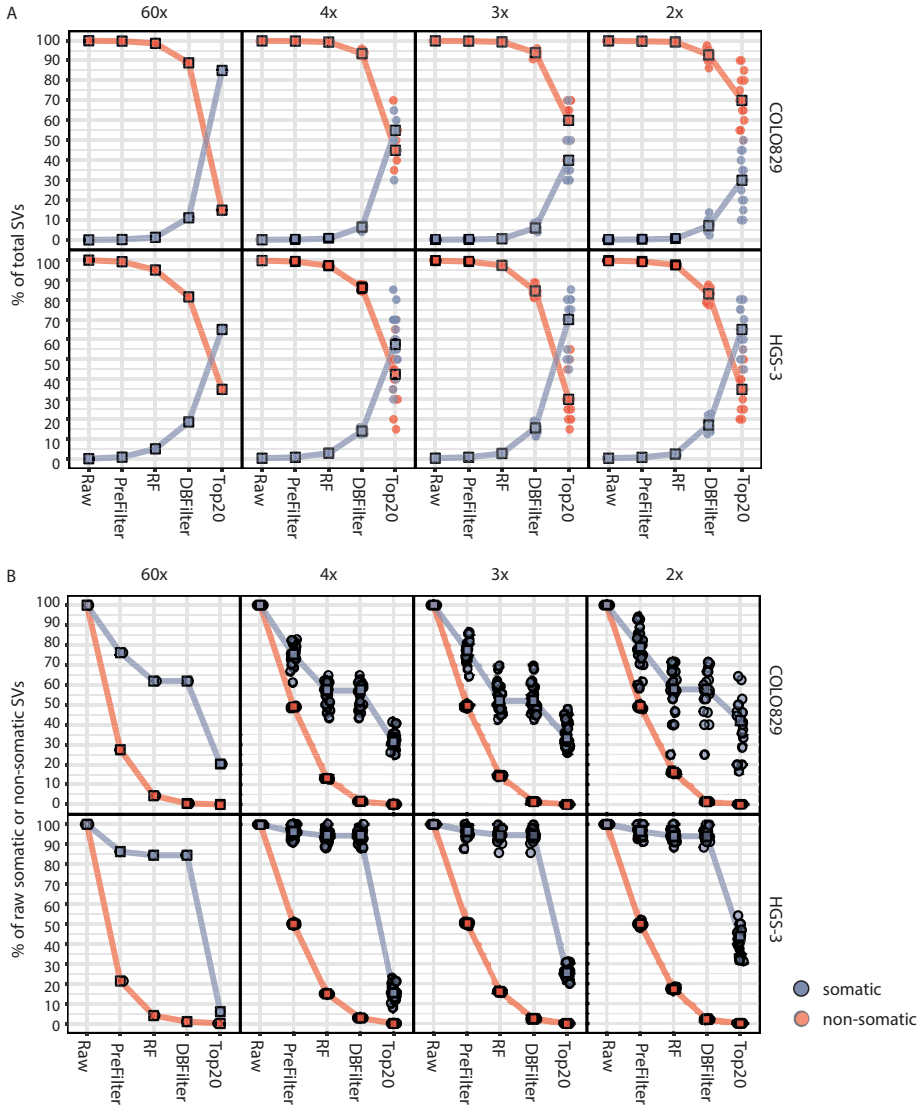
LIST OF SUPPLEMENTARY DATA

Figure S1.	Coverage and read length of COLO829 and HGS-3.
Figure S2.	Enrichment of somatic SV calls of COLO829 and HGS-3 after subsequent steps of the SHARC pipeline.
Figure S3.	Benchmarking of nanopore SV callers on low coverage nanopore sequencing data.
Figure S4.	Somatic vs germline SV length.
Figure S5.	Random forest performance on the Genome in a Bottle sample (GIAB).
Figure S6.	Nanopore sequencing time vs. somatic SV detection.
Figure S7.	Validation of somatic SV (sSV) of Ova2 biomarkers in cfDNA.
Figure S8.	Confirmation of presence of SVs in pre-amplified cfDNA.
Figure S9.	Technical aspects of dPCR.
Figure S10.	dPCR-based quantification of SVs in blood.
*Table S1.	Metrics of sequencing data
*Table S2.	Subsampling results of COLO829 and HGS3
*Table S3.	Random forest feature descriptions and Gini scores
*Table S4.	Breakpoint locations and probe design of Prostate cancer samples

**Table S1-4 are available online at: https://tinyurl.com/Ch5Suppl or scanning the QR code below*

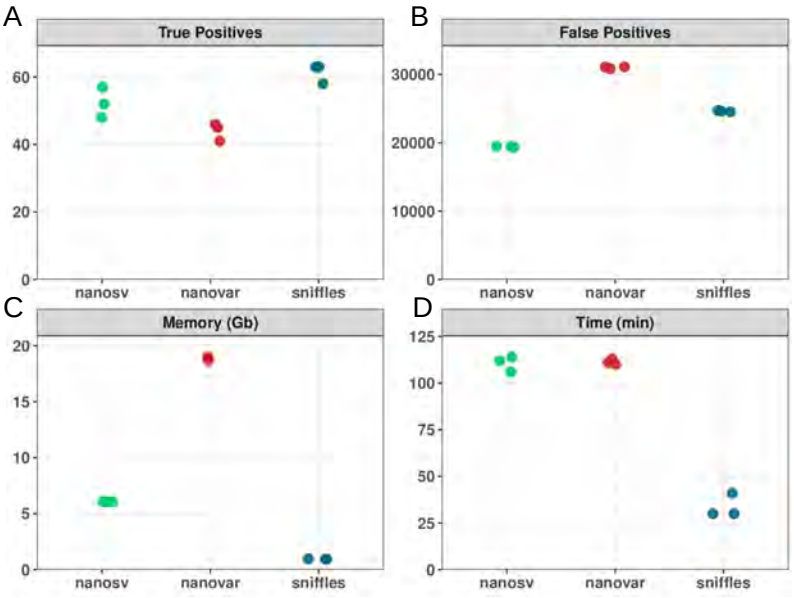


Supplementary Figure 1: Coverage and read length of COLO829 and HGS-3. Coverage **(A)** and read length **(B)** distribution for COLO829 and HGS-3 nanopore sequencing data. Dashed lines represent average.

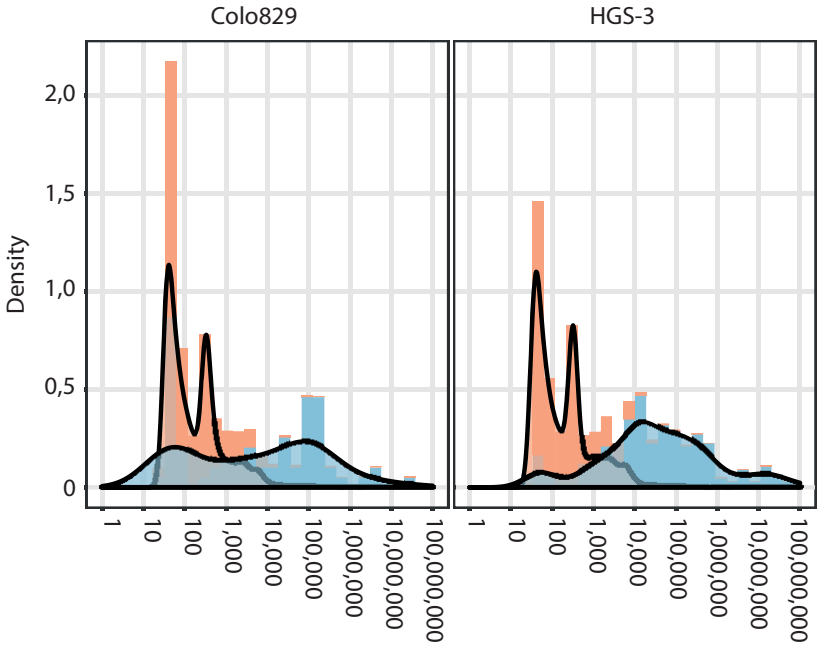


Supplementary Figure 2: Enrichment of somatic SV calls of COLO829 and HGS-3 after subsequent steps of the SHARC pipeline. The filtering steps were applied in a cumulative manner in the order displayed for subsampled Nanopore sequencing datasets from COLO829 and HGS-3. For each level of coverage, 20 independent subsampled datasets were generated and subjected to each filtering step in a cumulative manner. **(A)** Enrichment for somatic SVs after subsequent steps of the SHARC filtering pipeline. The blue and red lines/dots indicate the percentage of somatic and non-somatic SV calls after each filtering step of the pipeline for both COLO829 and HGS-3. The percentage of somatic and non-somatic SV calls is calculated relative to the sum of remaining somatic and non-somatic SV calls after each filtering step. Thus, 100% represents the total number of SV calls (somatic plus non-somatic) present at each step. **(B)** This figure panel is based on the same underlying data as for panel A, but here the percentage of somatic (blue) and non-somatic (red) SV calls is plotted relative to the total number of somatic and non-somatic SV calls detected at the first step, respectively. Thus, 100% represents the total number of non-somatic or somatic SV calls found initially in the raw data prior to filtering. While the percentage of non-somatic SV calls (red line/dots) decreases rapidly

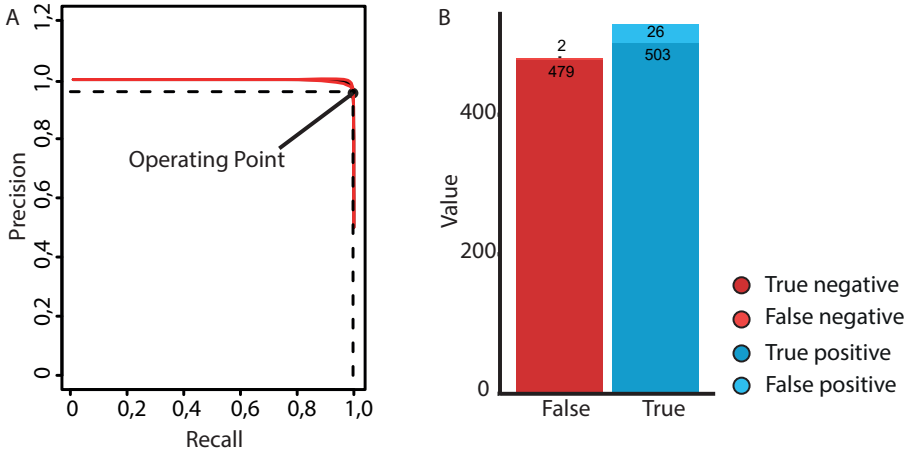
to very low percentages, the percentage of true positive somatic SV calls (blue line/dots) remains substantial (around 20%, depending on the sequence coverage). In low coverage subsets, all data points are shown and the square box represents the median value. RF: Random forest; DBFilter: Database filter.



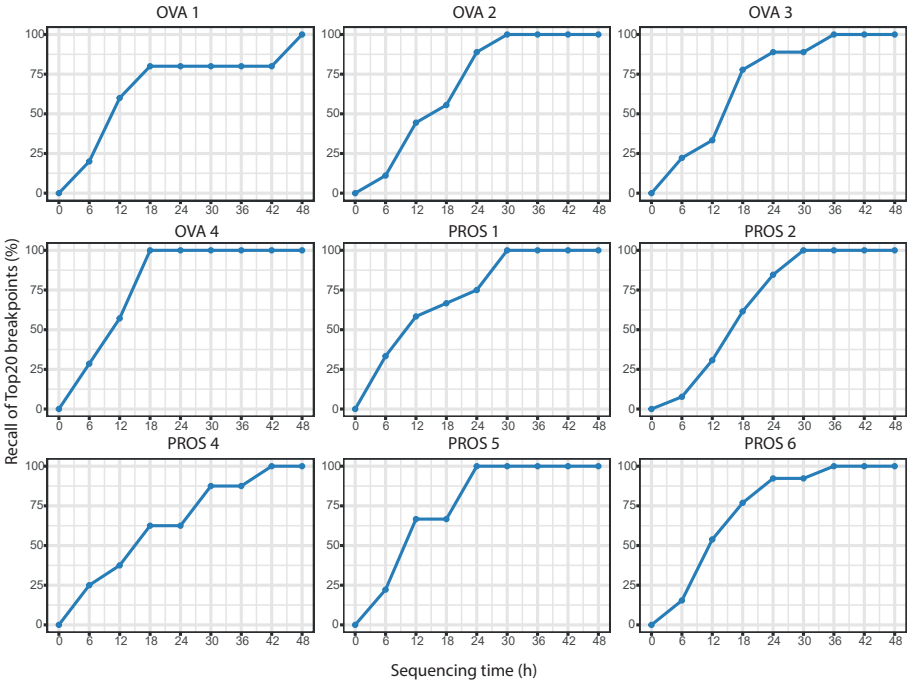
Supplementary Figure 3: Benchmarking of nanopore SV callers on low coverage nanopore sequencing data. The SV callers NanoSV, Sniffles and NanoVar are compared in terms of true positives (**A**), false positives (**B**) and required computation memory (**C**) and time (**D**). Triplicates of 5x randomly subsampled COLO829 data were used, and comparisons were performed against a short-read somatic SV reference set.



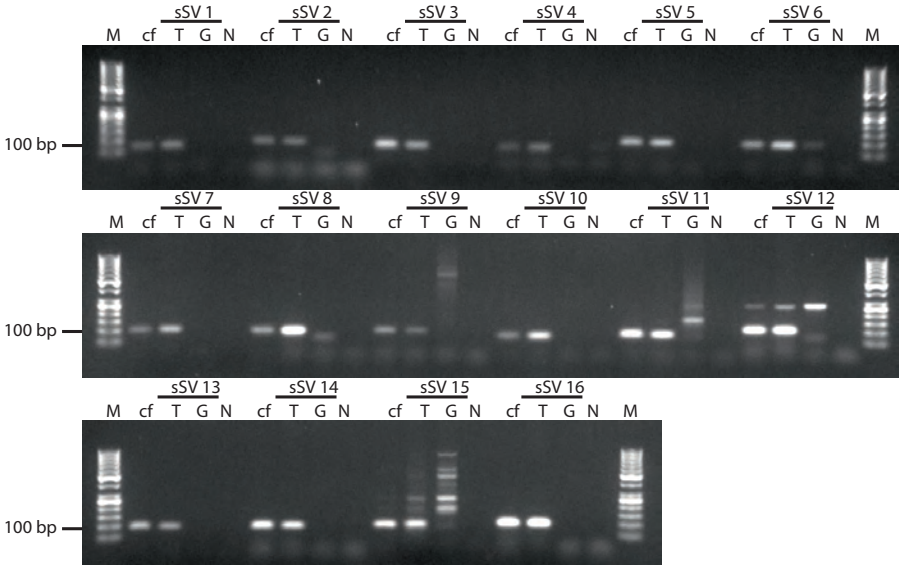
Supplementary Figure 4: Somatic vs germline SV length. Histogram and density plot of SV lengths of somatic and germline SVs from short-read data of COLO829 and HGS-3.



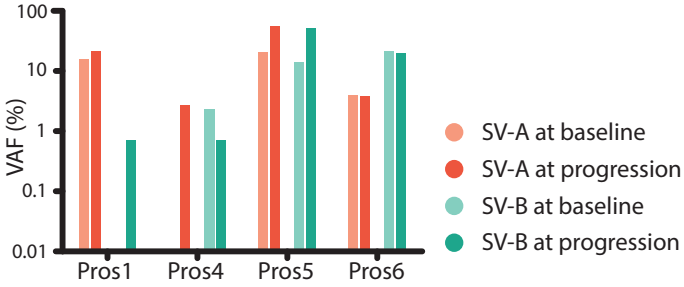
Supplementary Figure 5: Random forest performance on the Genome in a Bottle sample (GIAB). (A) Precision vs recall curve on the training set. Depicted is the operating point selected of 96% precision and 99.5% recall. **(B)** Random forest performance on the hold-out set.



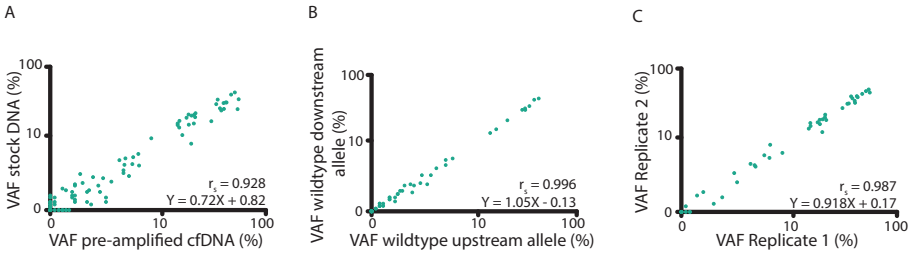
Supplementary Figure 6: Nanopore sequencing time vs. somatic SV detection. Plots showing the sequencing time and the recall of validated somatic SVs in 6-hour cumulative bins.



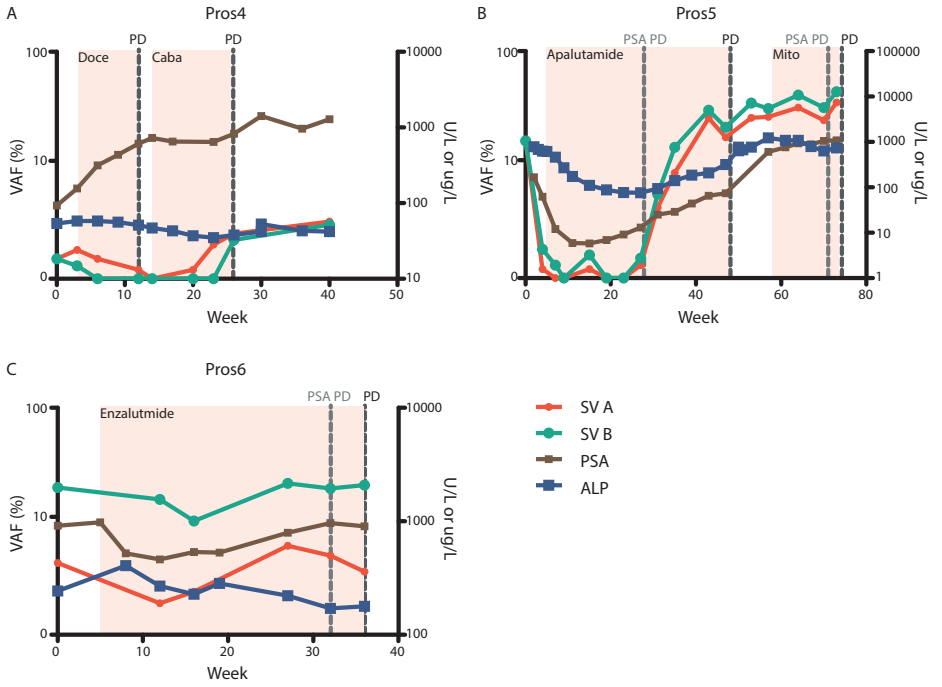
Supplementary Figure 7: Validation of somatic SV (sSV) of Ova2 biomarkers in cfDNA. sSVs of patient Ova2 were tested on cfDNA from ascites (cfDNA), tumor DNA (T), germline DNA (G) and water control (N). M = DNA ladder.



Supplementary Figure 8: Confirmation of presence of SVs in pre-amplified cfDNA. Detection of two patient-specific SVs in cfDNA from blood from four prostate cancer patients at baseline and at progression of disease with dPCR. Shown are VAFs.



Supplementary Figure 9: Technical aspects of dPCR. (A) Comparison of VAF in pre-amplified cfDNA and VAF in stock (non-pre-amplified) cfDNA. **(B)** Comparison of VAF based on up- and downstream wild type alleles. **(C)** Comparison of VAF in technical replicates of dPCR of pre-amplified cfDNA samples of Pros5 and Pros6.



Supplementary Figure 10: dPCR-based quantification of SVs in blood. Quantification of SVs in longitudinal cfDNA samples from blood in patient (**A**) Pros4, (**B**) Pros5 and (**C**) Pros6. In addition to VAFs of SVs, treatment, laboratory parameters (prostate specific membrane antigen (PSA), alkaline phosphatase (ALP)) and clinical progression of disease (PD) are visualized. Progression of disease based on a confirmed increase of prostate specific membrane antigen (PSA) of ≥25% above the nadir or baseline (PCWG3 criteria) was present in Pros5 and Pros6 (PSA PD). Doce, docetaxel; Caba, cabazitaxel; Mito, mitoxantrone.

AUTHOR CONTRIBUTIONS

WPK and JEV-I conceived the study. JEV-I, SdB and MJvR performed bioinformatic experiments and JEV-I, SdB, MJvR, CS and WPK analyzed the data. RJ and JEV-I packaged the SHARC pipeline into a GNU Guix package. IR performed nanopore sequencing. CS, IR, WPK and SdB performed wet lab experiments. CJdW, LFvD, MPL and ACdJ provided patient samples and clinical information. JCAH and AdJ performed cfDNA quantifications. JEV-I, CS, ACdJ, JCAH, LFvD, JWMM, MPHMJ, MPL and WPK interpreted data. JEV-I, CS and ACdJ wrote the manuscript. MPL and WPK edited the manuscript which was then reviewed and approved by all authors.

ACKNOWLEDGEMENTS

The authors thank the former Kloosterman group at the UMC Utrecht and the Medical Oncology Department in the Erasmus MC for critical input. We thank Job van Riet for help with the design of dPCR assays. We thank Oxford Nanopore Technologies for providing the research prototype high sensitivity flow cells and the Utrecht Sequencing Facility for the nanopore sequencing. We thank all patients for providing the clinical specimens to perform this study.

6

TARGETED LONG-READ SEQUENCING PROVIDES PERSONALIZED BIOMARKERS FOR MINIMAL RESIDUAL DISEASE TRACING IN PEDIATRIC LEUKEMIA

Targeted long-read sequencing provides personalized biomarkers for minimal residual disease tracing in pediatric leukemia

Christina Stangl, Jose Espejo Valle-Inclan, Liset Westera, Ivo Renkens, Karen Duran, Bastiaan B. J. Tops, Emile E. Voest, Glen R. Monroe[$], Gijs van Haaften[$]

[$]*corresponding authors*

Manuscript in preparation

ABSTRACT

Pediatric leukemias are aggressive tumors and therefore require immediate diagnosis, treatment initiation and appropriate biomarkers to facilitate minimal residual disease (MRD) tracing. The occurrence of genomic rearrangements such as fusion genes are hallmark events in these cancers and their unique genomic breakpoint junctions serve as an attractive MRD biomarker. Furthermore, leukemic cells derived from the lymphoid lineage (e.g. ALL and CLL) harbor clonal variable (V), diversity, (D), joining (J) rearrangements in the immunoglobulin (Ig) and/or T-cell receptors (TCR) which are widely used for PCR-based MRD tracing approaches. Timely detection of the exact genomic breakpoint junction of the fusion genes and/or the dominant V(D)J clones is therefore essential; however, this often presents as a race against the clock with current multi-step diagnostic processes. We here applied FUDGE -a targeted long-read nanopore sequencing assay- to a panel of ten *KMT2A* or *SIL-TAL1* fusion-positive acute lymphoid leukemia samples. We show that we can comprehensively target and obtain high coverage across the two fusion loci as well as the complex Ig and TCR loci with a single sequencing run. Within 48 hours we validate 86% of the diagnostically defined MRD targets for these patients. Furthermore, we identify an additional set of patient-specific rearrangements that could be used for MRD tracing. We here utilize FUDGE to detect unique genomic breakpoint junctions in pediatric lymphoid leukemia samples and present it as an attractive alternative to current multi-step biomarker identification assays for an increase in speed and detection sensitivity.

INTRODUCTION

Leukemias are the most common form of pediatric cancers and account for approximately 30% of childhood cancer incidences[375]. Leukemias develop from hematopoietic progenitors in the bone marrow or the thymus and cause an abnormal increase of white blood cells[376]. Depending on the type of affected lineage (myeloid or lymphoid) and the kinetics of the disease (acute or chronic), leukemias can be subdivided into acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), chronic myeloid leukemia (CML) and chronic lymphocytic leukemia (CLL). These subtypes can be further subdivided based on molecular abnormalities or immunophenotypic presentation and present with differing prognostic predictions[377,378]. While the overall survival rate of pediatric leukemias has drastically increased over the past decades, with 90% for ALL and 75% for AML, refractory or relapsed patients still face poor prognosis, therefore requiring accurate diagnosis at initial presentation as well as at recurrence[377,379]. Key to the management of leukemias is monitoring of minimal residual disease (MRD) – small traces of leukemic cells that can still be detected by sensitive methods like flow cytometry or PCR[380]. MRD is a strong prognostic factor widely used for risk group stratification and other clinical decision-making[381,382]. Molecular MRD monitoring relies on genetic markers that specifically identify the leukemic cells. Genetic abnormalities such as the unique breakpoint junctions created by genomic rearrangements are optimal biomarkers that reflect the disease burden[69,166].

A large proportion of leukemic subtypes are characterized by recurrent genomic rearrangements which may lead to the formation of fusion genes[383,31]. Fusion gene formation is a complex mechanism and the originating fusion gene configurations are extremely versatile. Whereas some events involve highly recurrent fusion gene configurations, such as the common *BCR-ABL1* fusion or the rare *SIL-TAL1* deletion rearrangement, others mainly involve a sole recurrent fusion partner gene[384–386]. *KMT2A* for example, is a promiscuous fusion partner gene and has been reported in >130 different fusion configurations[387,388]. Additionally, the genomic position where the break occurs within the involved genes is very variable - even for highly recurrent fusion gene configurations. However, this high level of versatility results in unique genomic breakpoint junctions, an acquired and patient-specific property of the tumor cells, and can therefore be exploited as a distinct, trackable biomarker.

Leukemias that arise from the lymphoid lineage mainly comprise a clonal outgrowth of single transformed lymphoblasts[389]. Lymphoblasts are immature lymphocytes that will become the effectors of the adaptive immune system, which recognizes millions or billions of different antigens thanks to a highly diverse repertoire of antigen recognition

molecules on B- and T- cells[390,391]. This antigen receptor repertoire is generated during their maturation by the unique process of the variable (V), diversity (D), joining (J) recombination, a somatic recombination at the genetic level during which different gene segments rearrange[392]. This process, taken together with random nucleotide insertions and deletions at the joint sites (junctional diversity), provides a highly diverse repertoire of immunoglobulins (Igs) in B-cells and T-cell receptors (TCRs) in T-cells[393,394]. The results at the genomic level are newly formed and unique genomic junctions which connect the rearranged V(D)J segments, and ultimately form the unique antigen binding domains of the antigen receptor proteins. Depending on the developmental stage at which lymphoblasts undergo malignant transformation, they usually have already undergone one or more successive VDJ recombination events and thus harbor one or more unique rearrangements in their DNA: fingerprints of the leukemic cells that can be exploited as molecular markers for MRD monitoring[395].

Biomarkers for MRD assays must be patient specific and are therefore often based on the unique properties of the underlying genomic rearrangement[396]. Fusion genes can easily be detected by breakpoint-spanning primers (i.e. one primer upstream and one primer downstream of the breakpoint junction), while Ig/TCR rearrangements require a breakpoint-specific primer (i.e. one primer sitting on the breakpoint junction) combined with a germline probe and primer to ensure increased specificity[396,397]. Quantifications are usually performed with PCR-based assays, and return information on the effectiveness of the treatment as well as tumor dynamics[396,398]. Furthermore, the tumor material can be obtained through liquid biopsies (i.e. blood) in a minimal invasive manner. Hence, not only MRD but also therapy response can be effectively assessed in a longitudinal fashion. Current diagnostic approaches to identify MRD targets mainly include PCR-based assays, targeted NGS or WGS/RNA-Seq[399,400]. Routine DNA-breakpoint PCRs have been developed for the Ig/TCR rearrangements, however, for some labs no standardized assays are available for versatile fusion events like *KMT2A* fusions or even the recurrent configurations such as the *SIL-TAL1* rearrangement. For the majority of recurrent fusion events transcriptomic breakpoints are utilized as the breakpoint detection is simplified through its restriction to exon-exon junctions and predeveloped probes can ensure high quality detection and quantification assays[401]. While a transcriptomic breakpoint may be less representative of the actual number of leukemic cells, their simple and reliable methodology and rapid turnaround time has made them a standardized diagnostic assay. (Targeted) NGS approaches are now being implemented in diagnostic labs due to their mostly unbiased ability to detect fusion genes at the RNA or DNA level, however, these come with more complex analysis strategies, validation requirements and longer turnaround times[399]. Furthermore, RNA-Seq is not suited to detect promoter fusions or accurately define the genomic breakpoint location[402]. Optimally, a diagnostic assay

would comprehensively identify genomic MRD targets from fusion genes and Ig/TCR loci, in a rapid (e.g. less than one week), comprehensive and straightforward workflow with minimal hands-on time requirements.

We recently developed FUDGE -a targeted and directional nanopore sequencing assay- to rapidly identify recurrent fusion genes with unknown fusion gene partners and extract the exact genomic breakpoint location[30]. Based on this strategy, we developed a nanopore sequencing assay ideally suited for the detection of personalized genomic MRD targets from gene fusions and Ig/TCR rearrangements in lymphoid leukemias. We retrospectively tested a panel of five *KMT2A* positive B-cell acute lymphoblastic leukemias (B-ALL) and five *SIL-TAL1* positive T-cell acute lymphoblastic leukemia (T-ALL) samples. We show the capability of the assay to comprehensively target fusion genes as well as the Ig and TCR loci and to detect the sequence of genomic breakpoints for use in the subsequent design of MRD assays. We confirm 86% of the MRD targets that were diagnostically determined for these patients and identify an additional 32 potential patient-specific MRD targets - all within 48 hours.

6

RESULTS

KMT2A AND SIL-TALI ASSAY DESIGN

We previously developed FUDGE[403], a targeted and directional nanopore sequencing assay, to identify recurrent fusion genes irrespective of fusion partner and genomic breakpoint position. We here apply FUDGE to a panel of KMT2A and SIL-TAL1 fusion-positive B-ALL and T-ALL patient samples, respectively, to identify unique genomic breakpoint junctions which may serve as suitable biomarkers for subsequent quantitative PCR-based MRD testing with breakpoint-spanning primers (**Figure 1A**). FUDGE couples target-selected and strand-specific CRISPR/Cas9 activity, which facilitates enrichment by selective sequencing adapter ligation, to long-read real-time nanopore sequencing (**Supplementary Figure 1**). KMT2A, a recurrent fusion partner gene in leukemias, has been reported with >130 different fusion gene configurations. Furthermore, there are two clusters within the KMT2A gene spanning approximatively 12.4 kb where previous breakpoints have been reported[387]. Therefore, we designed two crRNAs (crK-e7 and crK-e19) targeting these two main breakpoint clusters (**Figure 1B**). SIL-TAL1, a specific and recurrent rearrangement between SIL and TAL1 adjacently located on chromosome 1, is induced through a site-specific deletion between the genes of approximately 85 kb or chromosomal translocations in approximately 20% of T-ALL patients[404]. To identify the exact genomic breakpoint location, we designed two crRNAs

to guide sequencing to the 7.7 kb region within TAL1 and one crRNA to cover the 3.7 kb region within SIL (**Figure 1C**).

KMT2A AND SIL-TALI DETECTION

To validate our assay design, we sequenced ten acute lymphoblastic leukemia samples: five B-ALL (B-ALL1, B-ALL2, B-ALL3, B-ALL4, and B-ALL5) and five T-ALL (T-ALL1, T-ALL2, T-ALL3, T-ALL4, and T-ALL5). For the B-ALL1 and B-ALL2 samples, previous diagnostic efforts identified KMT2A-MLLT1 fusions, however, for B-ALL1 the fusion could only be detected by RNA-sequencing, but could not be validated by PCR. For samples B-ALL3, B-ALL4, and B-ALL5 we performed a blinded analysis and did not know the fusion partner before our analysis. For all five T-ALL samples SIL-TAL1 rearrangements were detected by heteroduplex analysis.

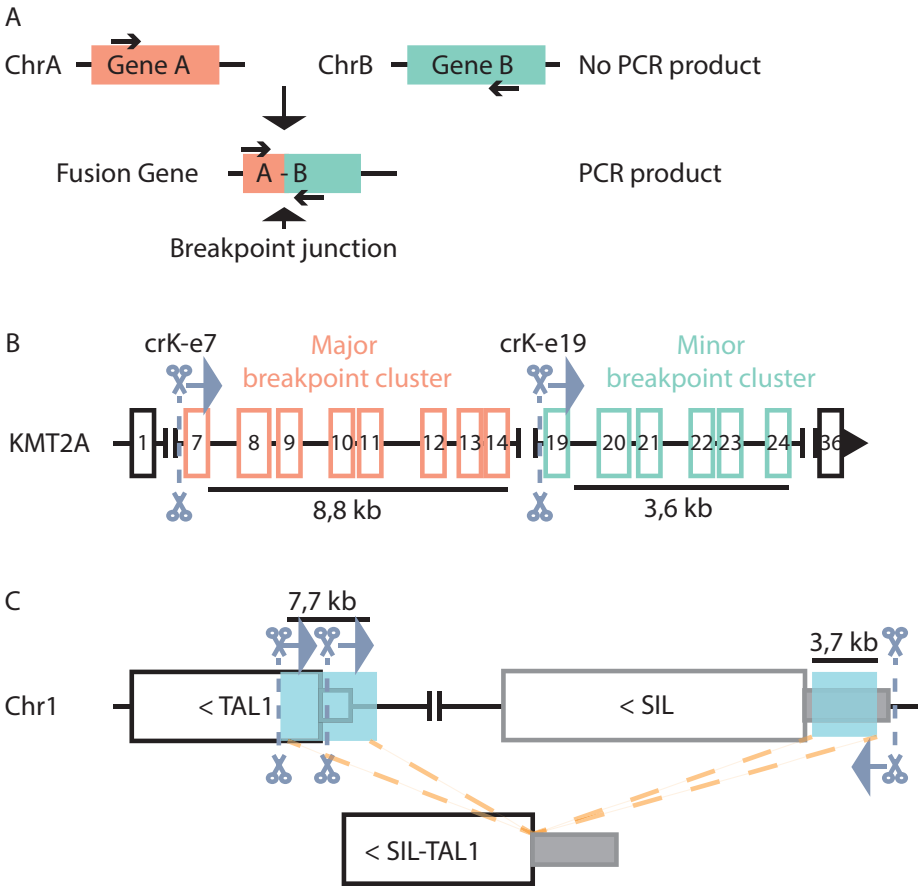


Figure 1. Schematic outline of fusion genes and the respective crRNA design (**A**) Schematic outline of fusion gene formation. Gene A and Gene B are fused together and provide a unique breakpoint

junction. PCR primers are designed on both sides of the fusion. A PCR with breakpoint-spanning primers will only result in a product for the rearrangement. (**B**) Schematic outline of *KMT2A* with its two breakpoint clusters (major: red; minor: green). Scissors indicate where the crRNAs target Cas9 to cut and arrows indicate the desired sequencing direction. (**C**) Schematic outline of *SIL* and *TAL1*. Common breakpoint areas are highlighted by a blue box. Scissors indicate where the crRNAs target Cas9 to cut and arrows indicate the desired sequencing direction. Genomic regions are not scaled.

We performed target enrichment for each sample with appropriate sets of crRNAs (**Supplementary Table 1**) and sequenced each sample on one Oxford Nanopore Technologies MinION flow cell (R9.4). After mapping to the reference genome (GRCh38), the genome wide coverage per sample was 0.028x (B-ALL1), 0.04x (B-ALL2), 0.10x (B-ALL3), 0.54x (B-ALL4), 0.67x (B-ALL5), 0.22x (T-ALL1), 0.17x (T-ALL2), 0.032x (T-ALL3), 0.12x (T-ALL4) and 0.037x (T-ALL5) (**Supplementary Table 1**). However, due to the targeted sequencing approach, on-target coverages increased to 2-352x for the KMT2A locus in the B-ALL samples and 47-290x at the *SIL* and *TAL1* loci in the T-ALL samples. (**Figure 2A and Supplementary Table 1**).

To detect *KMT2A* fusion genes we applied NanoFG, which we previously developed to identify fusion genes from nanopore sequencing data[403]. NanoFG identified, as expected, a *KMT2A-MLLT1* fusion gene with two fusion-supporting reads in the B-ALL2 sample (**Supplementary Table 2**). For the blinded samples B-ALL3 and B-ALL5 we identified a *KMT2A-MLLT2* fusion with four and 31 fusion-supporting reads respectively, which was later confirmed by comparison with outsourced results obtained from a laboratory specializing in KMT2A partner detection. In addition to this fusion result, a reciprocal *MLLT2-KMT2A* fusion was detected in the B-ALL5 sample with 28 fusion-supporting reads that was not previously detected (**Supplementary Figure 2, Supplementary Table 2**). For B-ALL4, the diagnostic laboratory identified a *KMT2A-MLLT2* fusion, however, neither manual investigation in IGV genome browser[405] nor NanoFG could confirm this fusion within our targeted sequencing data. In accordance with diagnostic efforts, NanoFG did not identify genomic evidence of the *KMT2A* fusion for the B-ALL1 sample. We manually confirmed the lack of fusion-spanning reads within the *KMT2A* and *MLLT1* genes with IGV. NanoFG is specifically designed to detect fusion genes with breakpoints within both of the involved fusion partners.

Therefore, NanoFG did not call any of the *SIL-TAL1* rearrangements within the T-ALL samples as the *SIL-TAL1* rearrangement may be induced through breakpoints outside of annotated genes (including promoter, both UTRs, and exonic/intronic regions). Therefore, we performed SV calling with Sniffles[342] within the *SIL* and *TAL1* loci. We then identified all five *SIL-TAL1* rearrangements with 238x (T-ALL1), 145x (T-ALL2), 50x (T-ALL3), 177x (T-ALL4) and 49x (T-ALL5) breakpoint-spanning reads (**Figure 2B, Supplementary Table 2**). While the breakpoints within *SIL* were located within the 5'UTR of the gene, all breakpoints within *TAL1* were located about 185 bps upstream

of the *TAL1* gene (**Supplementary Table 2**), therefore restricting the detection by NanoFG. Hence, we show that we are able to reliably identify and confirm fusion genes and genomic rearrangements with a targeted sequencing approach.

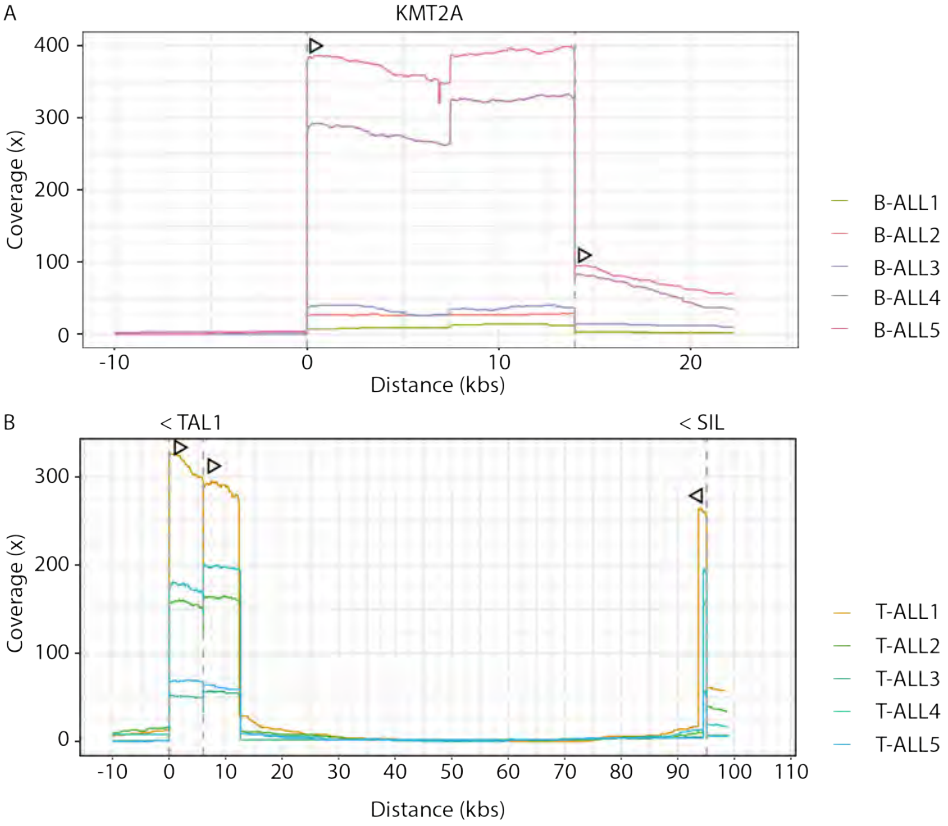


Figure 2. KMT2A and SIL-TAL1 coverage plots (A) Coverage plot for the samples B-ALL1, B-ALL2, B-ALL3, B-ALL4, and B-ALL5 between the cut-sites in KMT2A **(B)** Coverage plot for the samples T-ALL1, T-ALL2, T-ALL3, T-ALL4, and T-ALL5 between the cut-sites in TAL1 and SIL. Dotted lines (grey) show cut-sites and arrows indicate the desired sequencing direction.

IG AND TCR LOCI ASSAY DESIGN

Lymphocytic leukemias are clonal amplifications of cells derived from the lymphoid lineage, and therefore harbor unique and clonal V(D)J rearrangements. We exploited this feature to detect additional MRD targets. For this reason, in addition to targeting the *KMT2A* and *SIL/TAL1* loci, we comprehensively targeted the Ig (IGH, IGK, IGL) and TCR (TRA, TRB, TRD, TRG) loci. Each locus harbors a specific number of V, (D), and J segments (except IGK, IGL, TRA, and TRG which do not contain D segments)[406]. These segments, which themselves are usually <300 bps, are distributed over large genom-

ic regions (range: 130 kb - 1.4 Mb). However, during genomic somatic recombination, one V, (D), and J segment are juxtaposed, forming a unique rearrangement comprising a random amount of nucleotide insertions and deletions at the junction (**Figure 3A**). Post-recombination, these V(D)J rearrangements not only encompass a much smaller genomic region through the removal of intermediate gene segments but also form specific breakpoint junctions which can be targeted by breakpoint-specific PCR assays (**Figure 3B**)[396]. We aimed to identify the unique breakpoint junctions within these loci as well as the predominant clonal V(D)J rearrangements per patient by covering the genomic area from the last J segment until the first D/V segment (**Figure 3A**). To accomplish this, crRNAs selectively directing sequencing reads from the last J segment towards the first D/V segment were designed (**Figure 3C-3I**). Since some of these genomic areas are large (up to 70 kbs between the first and last J segment), sequencing reads (average ~10 kb) originating from a cut adjacent to the last J segment will not reach the first D/V segment. For these instances we applied a tiled crRNA approach[403], targeting these loci every ~8 kb, facilitating uniform coverage across the targeted loci (**Figure 3E-I**). Furthermore, the IGK locus can undergo consecutive rearrangements involving the intron RSS and Kdeletion (Kdel) segments[407]. Depending on the state of rearrangement, the Kdel is too far from the J segments (**Figure 3D**, IGK(1)) or within reach (**Figure 3D,** IGK(2)), requiring the design of two separate crRNAs.

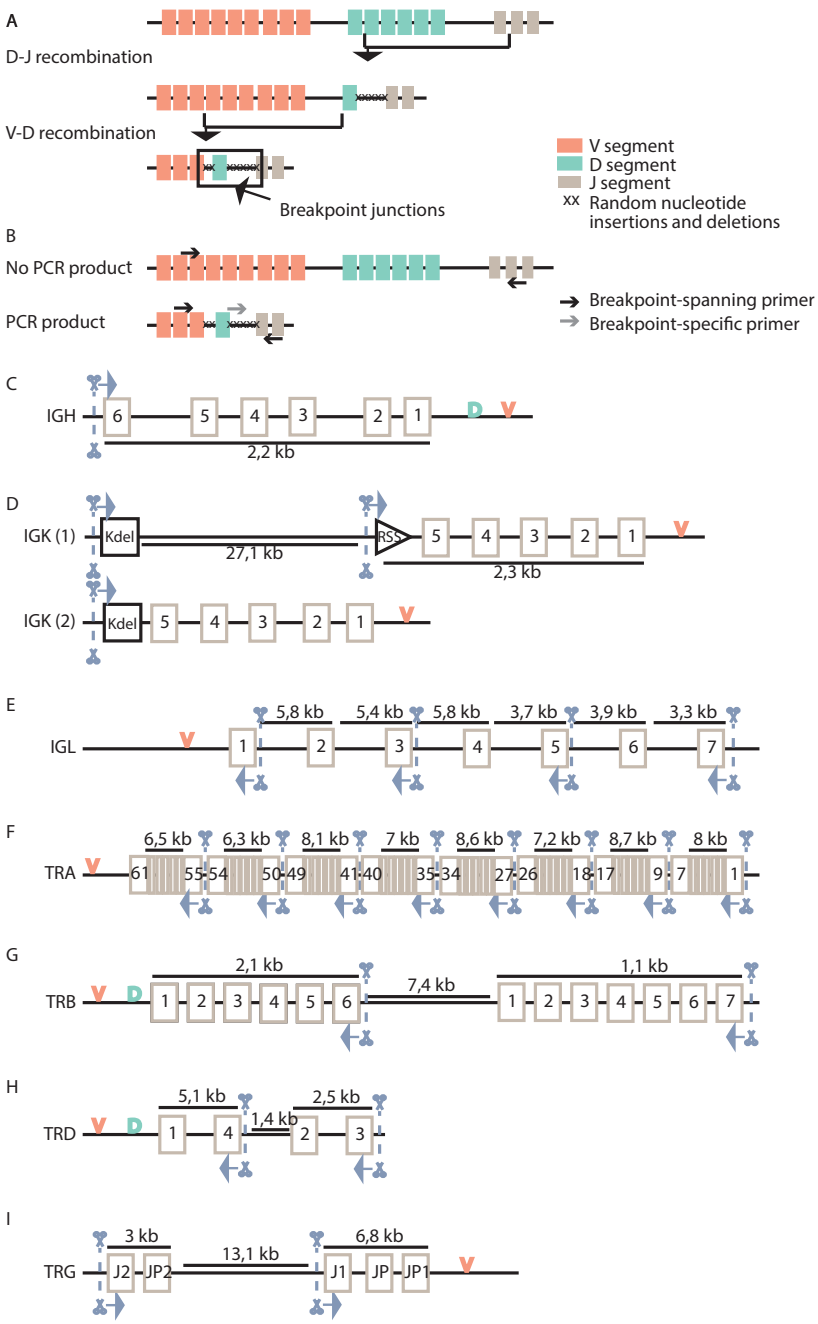


Figure 3. Schematic outline of VDJ recombinations and crRNA design (figure on previous page). (A) Schematic outline of V(D)J recombinations within the IGH locus. First, one D (green) and one J (beige) segment get juxtaposed, while the intersecting genomic region gets excised and random nucleotide insertions and deletions (xx) occur. Then, one V (orange) and one D segment get juxtaposed, while the intersecting genomic region gets excised and random nucleotide insertions and deletions occur. For the

IGK, IGL, TRA, and TRG loci, only V to J recombinations take place as these lack D segments. The evolved V(D)J configuration including the random nucleotide insertions and deletions presents specific genomic breakpoint junctions. **(B)** Primers spanning the newly formed breakpoint junction (black) or specific to the breakpoint junction (grey) can be designed and will only yield a positive PCR result in the case of a rearrangement. Schematic depiction of the **(C)** IGH **(D)** two possible genomic events of IGK, **(E)** IGL, **(F)** TRA, **(G)** TRB, **(H)** TRD, and **(I)** TRG. J gene segments are boxed. Scissors indicate where the crRNAs target Cas9 to cut and arrows indicate the desired sequencing direction. Genomic regions are not scaled.

To validate that we can perform targeted sequencing of multiple Ig and TCR loci, we targeted the ten leukemia samples - in addition to the *KMT2A* and *SIL-TAL1* loci - simultaneously with crRNAs for the Ig and TCR loci (**Figure 3C-I and Supplementary Table 1**). B-ALL1, B-ALL2 and T-ALL1 were targeted with the fusion crRNAs as well as the IGH, IGK, IGL, TRB and TRG crRNAs (**Supplementary Table 1**). B-ALL3, B-ALL4, B-ALL5, T-ALL2, T-ALL3, T-ALL4 and T-ALL5 were additionally targeted with crRNAs for the TRA and TRD loci, a total of 25 crRNAs per sample, providing examples of parallel fusion and comprehensive Ig/TCR sequencing.

The average per-base coverage for all samples at the different Ig and TCR loci was 0.83x (IGH), 0.62x (IGK), 5.97x (IGL), 10.4x (TRA), 3.39x (TRB), 2.77x (TRD), and 10.60x (TRG) (**Supplementary Table 1**). We observed a sharp increase of coverage at the cut sites (**Figure 4**). For the IGH, IGK and TRG loci, where we only target with one crRNA adjacent to the last J segment or the crRNAs were spaced far apart, we observe reads longer than 20 kb (**Figure 4A-B and F**). Noteworthy, the utilized DNA was not specifically isolated for long-read sequencing approaches, and thus, these read lengths could be pushed even longer by applying appropriate extraction methods. For the IGK, IGL, TRA, TRB, TRD and TRG we clearly observe the effectiveness of the tiled crRNA approach to cover the entire loci (**Figure 4B-F**). For TRA, we are able to span a 60 kb window with eight consecutive crRNAs (**Figure 4D**). Visual investigation of the coverage plots already identifies large scale rearrangements events, e.g. by the observed drop in coverage in between crRNA cut-sites (**Figure 4E-F**). The average coverage at the cut-sites in the Ig and TCR loci varies from 19x to 230x between the different loci (**Supplementary Table 1**), which could be due to underlying biological factors or the efficiency of the crRNAs. We also observe differences in coverages at the cut sites within one locus (e.g. TRA), which would support a theory that crRNA efficiency drastically influences the enrichment capabilities. However, investigating the relationship between coverage at the cut-site and on/off-target scores of the crRNAs as provided by IDT[408], doesn't show an obvious influence of crRNA efficiency on the coverage (**Supplementary Figure 3A and 3B**). We do observe a correlation between sequencing throughput and coverage at the cut sites, indicating that a minimal sequencing throughput may be required for a successful enrichment (**Supplementary Figure 3C**).

Taken together, these data show that we are able to comprehensively target fusion genes and Ig as well as TCR loci in one MinION sequencing run. With the tiled crRNA design

we achieve enriched sequencing coverage across large areas allowing a comprehensive investigation of genomic rearrangements within these loci.

MRD TARGET DETECTION

To validate that we can identify unique genomic breakpoint junctions from the enriched nanopore sequencing data, we collected diagnostic information about the 36 MRD targets there were utilized for these patients (**Table 1**). These targets included both gene fusions and Ig/TCR rearrangements. We implemented a bioinformatic pipeline to automatically detect breakpoint junctions within the fused genes as well as the Ig/TCR loci. As the detection of the *KMT2A* and *SIL-TAL1* fusion is already provided by NanoFG and Sniffles, respectively, we focussed on the detection of MRD targets within the Ig and TCR loci. For this, we extracted all reads falling within the genomic loci of the Ig and TCR loci and performed SV calling independently for each locus. We performed breakpoint-spanning PCRs on the detected breakpoints, using the corresponding tumor sample and a polyclonal healthy donor as control sample. This resulted in a PCR product in 97 out of the 140 assays (69%), highlighting the ability of our approach to identify Ig and TCR rearrangements from enriched nanopore sequencing data. However, we observed a PCR product for the control sample as well as the tumor sample in the majority of the cases. This was to be expected as we perform, in this step, an endpoint PCR assay without the use of a breakpoint-specific probe to select patient-specific clonal rearrangements. Furthermore, the control sample consists of a mixture of several healthy donor buffy coats, providing a rich background of possible rearrangements. In 38 (39%) of these cases, however, the amplification band in the tumor sample displayed a distinct pattern compared to the control sample, which presented a shifted size, lower intensity or smear, indicating a higher specificity of the primer to the tumor rearrangement (**Supplemental Figure 4**).

We then annotated the resulting SVs with the genomic location of the different Ig/TCR genes. Through this approach, we found evidence of 31/36 (86.1%) of the MRD targets used in diagnostics for these patients (**Table 1**). For the five remaining targets, two lacked any evidence for the rearrangement while two other targets had only a single read supporting the target, insufficient for SV detection (**Table 1**). Interestingly, for the remaining rearrangement, diagnostic documentation reported a DH3.10-JH2 rearrangement in the B-ALL5 sample. Our long-read sequencing data did not provide proof of this specific connection, however, instead reporting multiple connections between DH3.10 and JH4 as well as an inversion between JH4 and JH2 with the same supporting reads, indicating a more complex rearrangement than originally assumed by short-read NGS (**Supplementary Figure 5**). In addition to the MRD targets already used by diagnostics, we were

able to detect a total of 20 supplementary breakpoints that connect V, D or J segments in the Ig or TR loci, which represent potential MRD targets. Of these, 8 belong to the B-ALL5 sample, 4 to the T-ALL1 sample, 3 to the B-ALL4 sample, 2 to the B-ALL2, 2 to the T-ALL4 sample and 1 to the T-ALL3 sample (**Table 1, Supplementary Table 3**). Furthermore, there are a total of 12 validated SVs that, even though they do not connect V(D)J segments, could be used as patient-specific biomarkers with specific break-junction assays in a similar fashion as previous studies[69,409].

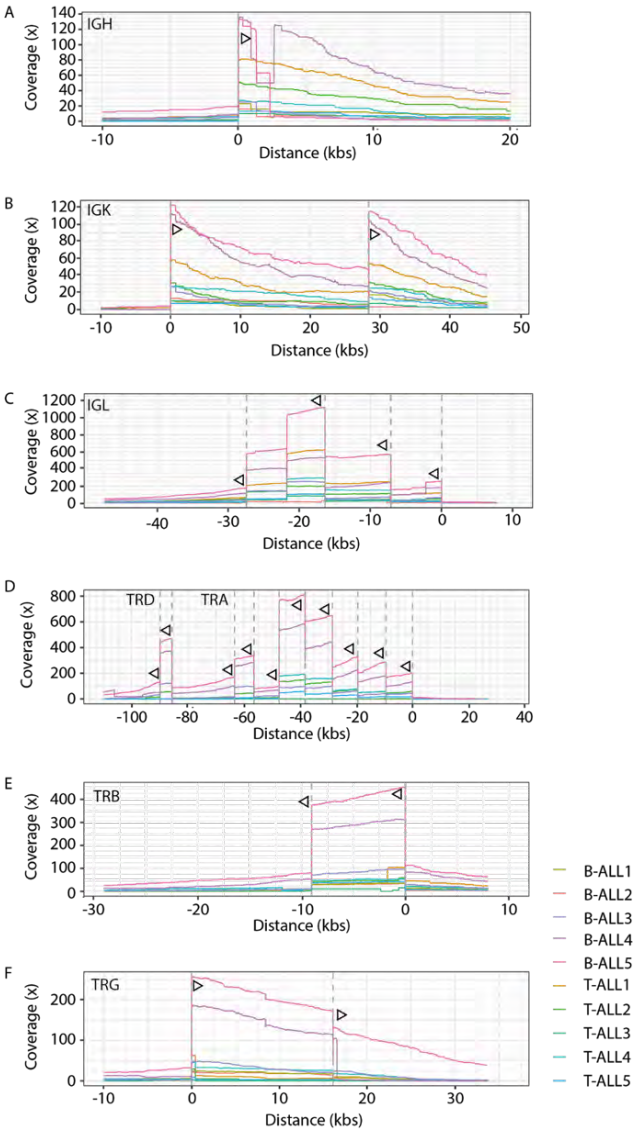


Figure 4. Ig and TCR coverage plots: Coverage plots for the samples B-ALL1, B-ALL2, B-ALL3, B-ALL4, B-ALL5, T-ALL1, T-ALL2, T-ALL3, T-ALL4, and T-ALL5 for the **(A)** IGH, **(B)** IGK, **(C)** IGL, **(D)** TRA and

TRD, (**E**) TRB, and (**F**) TRG loci. Dotted lines show cut-sites and arrows indicate the sequencing direction.

To further validate the rearrangements detected with our method, we performed Sanger sequencing of 16 selected rearrangements based on one of the following criteria: diagnostic Ig/TCR MRD target, read support ≥ 10, breakpoints connecting two V,D or J segments or a clear difference in breakpoint-PCR between the tumor and control sample. In parallel, we performed consensus calling on the nanopore reads to obtain the accurate sequence at the breakpoint junctions. In 8 cases (50%), consensus calling was not successful, of which 6 could be traced back to insufficiently low read support. In the other half where we obtained consensus, in 2 cases (25%) of cases we obtained a single consensus sequence whereas in 75% (6/8) of the cases, multiple consensus sequences were generated. When aligning the results from the Sanger sequencing to the nanopore consensus reads we observed a 100% match in 31% (4/13), one mismatch in 23%, (3/13) and two or more mismatches in 46% (6/13) of the cases. Taken together, this data shows that our targeted sequencing approach of Ig and TCR loci validated 86% of previously validated MRD targets and identifies an additional set of 32 PCR-validated rearrangements for potential MRD tracing in this cohort. However, verification of the exact sequence at the breakpoint by Sanger sequencing is still a requirement, especially in cases where rearrangements show a low read support by nanopore sequencing.

The time window from the start of the sequencing library preparation to obtaining the raw data was a maximum of 48 hours. Subsequently, bioinformatic detection of fusion genes and genomic breakpoints of Ig and TCR rearrangements including mapping, SV calling and automated primer design was performed in less than 3 hours. Thereafter, PCR assays and subsequent Sanger sequencing to obtain the accurate sequence at the breakpoint junction for junction-specific primer design are necessary, which in our hands could be performed in 72 hours until sequences were obtained. Hence, depending on the speed of the primer delivery, sequencing data for subsequent MRD assay design could be ready available after five days of sample obtention.

Table 1. Concordance of MRD targets identified by diagnostics and our assay

SAMPLE	DIAGNOSTIC TARGET	DETECTED	COMMENT	ADDITIONAL VDJ TARGETS	ADDITIONAL REARRANGEMENTS
B-ALL1	DH7.27-JH5b	yes	.	0	2
	VH6-1-JH2	yes	.		
B-ALL2	Dh1.26-Jh5b	no	only 1 read	2	0
	DH4-23-Jh4b	yes	.		
	KMT2A-MLLT1	yes			

SAMPLE	DIAGNOSTIC TARGET	DETECTED	COMMENT	ADDITIONAL VDJ TARGETS	ADDITIONAL REARRANGEMENTS
B-ALL3	Dh3-9-Jh4b	yes	.	0	1
	KMT2A-AFF1	yes	.		
B-ALL4	Dd2-Dd3	yes	.	3	1
	Vg11-Jg1.3/2.3	yes	.		
B-ALL5	KMT2A-MLLT2	yes	.	8	2
	DH2-8*02-JH4b	yes	.		
	Vd2-Dd3	yes	.		
	Dh3-10-Jh2	no*	connection between Dh3.10-Jh4 and Jh4-Jh2 with shared supporting reads		
T-ALL1	Vg2-Jg1	no	no	4	4
	Vb12-Jb1.2	yes	.		
	Vb6-4-Jb2.1	yes	.		
T-ALL2	SIL-TAL1	yes	.	0	1
	Vd1-Dd3-Jd1	yes*	V1-J1 evidence, jumping the Dd3 segment		
	Vg8-Jg2.3	yes	.		
	Vg10-Jg2.3	yes	.		
	Vb20.1-Db2	yes*	V20.1-J2.7 evidence, jumping the Db2 segment		
	Db1-Jb1.4	yes	.		
T-ALL3	SIL-TAL1	yes	.	1	1
	Vg4-Jg2.3	no	only 1 read		
	Vb20-Db2-Jb2.3	yes*	V20.1-J2.3 evidence, jumping the Db2 segment		
	Db2-Jb2.7	yes	.		
T-ALL4	SIL-TAL1	yes	.	2	0
	Db2-Jb2.1	yes	.		
	Vg8-Jg1.1	yes	.		
	Vg2-Jg2.3	yes	.		
	Vb15-Db1-Jb1.1	yes*	V15-J1.1 evidence, jumping the Db1 segment.		
T-ALL5	SIL-TAL1	yes	.	0	0
	Db2-Jb2.3	yes	.		
	Vg4-Jg2.3	yes	.		
	Vb29.1-Db1-Jb1.1	no	no evidence		
	Db1-Jb1.4	yes	.		
Total	36	31 (86.1%)		20	12

DISCUSSION

Timely detection of patient and leukemia-specific biomarkers to monitor treatment response and MRD is key for pediatric leukemia patients[380,396]. Currently, these biomarkers mainly entail transcriptomic breakpoint junctions for recurrent fusion genes or genomic breakpoint junctions for Ig and TCR rearrangements[395,401]. While the identification of transcriptomic breakpoint junctions has been long established and widely used for MRD targets, the use of e.g. cDNA-based biomarkers is suboptimal for accurate assessment of tumor cell quantifications. Furthermore, current techniques to detect genomic breakpoint junctions, especially for the complex Ig and TCR rearrangements, are laborious, and often involve multi-step protocols[410]. Following the detection of appropriate MRD targets, patient-specific PCR assays need to be developed, another time-consuming effort which may encounter unexpected setbacks such as poor primer characteristics[398]. Hence, the faster an initial MRD target can be detected, the more time is available to design reliable patient-specific quantification assays. To facilitate this, we developed a targeted nanopore sequencing assay, which provides comprehensive sequencing coverage of preselected fusion gene loci as well as the complex Ig and TCR loci within 48 hours. Furthermore, this data allows rapid identification of genomic breakpoint junctions within these loci to be utilized as personalized MRD targets.

We retrospectively applied the assay to five *KMT2A* rearranged B-ALL and five *SIL-TAL1* positive T-ALL samples. For eight of the samples we could report the genomic breakpoint coordinates of the fusion genes and design breakpoint-spanning primers within 48 hours of sequencing library preparation. Upon manual investigation as well as bioinformatic analysis of the sequencing data, we found proof of 86% of the diagnostically defined MRD targets for these ten patients, missing only four targets in total. For the DH3-10-JH2 rearrangement in B-ALL5, we could not find direct evidence, however, our long read sequencing data suggests a more complex rearrangement involving the DH3-10, JH2 and JH4 segments. Additionally, unbiased and independent investigation of potential MRD targets with the SV caller Sniffles focussed on the Ig and TCR loci yielded in total 32 additional validated rearrangements, which represent potential MRD candidates. Most of our detected rearrangements yielded a positive PCR product, albeit most were not patient-specific and also showed an amplification band in the polyclonal healthy control. This highlights the ability to detect Ig and TCR rearrangements from enriched nanopore sequencing data.

We here for the first time performed targeted genomic long-read nanopore sequencing on the complex human Ig and TCR loci. These loci, which are very large in their germline state, harbour an extensive amount of V, (D), and J segments[411]. After random joining

of one V, (D), and J segment, with the inclusion and deletion of random nucleotides in the joint regions, and excision of intermediate gene segments, these newly formed rearrangements comprise a small but highly unique genomic region of less than 1kb[394]. Furthermore, the individual gene segments (e.g. J segments) show high homology to one another which complicates the design of specific crRNAs and increases the chance for off-target effects. A recent study performed target enrichment capture using probes on the mRNA transcripts of these regions to track transcriptome profiles of clonal lymphocyte populations in a breast cancer patient[412], however, no systematic targeting of these entire genomic loci was performed. With our approach we targeted the respective genomic loci from the last J segment, directing reads towards the D or V regions. In most cases, we had to apply a tiled crRNA approach to uniformly span this whole stretch. In total we designed 20 crRNAs to comprehensively cover all J regions of the IGH, IGL, IGK, TRA, TRB, TRD and TRG. While we do see enrichment at all of the targeted cut-points, we observe varying efficiency of the enrichment within and in between loci. Even though we do not observe a clear correlation between predicted crRNA efficiency and enrichment, intra-locus differences likely point to varying efficiency of the crRNAs. Optimized crRNA designs and prior *in vitro* testing of the crRNA efficiency could help to achieve a maximal enrichment across all target sites[413].

The accuracy of nanopore sequencing at the single nucleotide level is currently suboptimal which may lead to difficulties in identifying the exact junctional sequence of V(D)J rearrangements or with the classification of these high homology gene segments[414,415]. Some of these segments are only 10 bps in size or differ by only one nucleotide, which may further complicate accurate assignment of a unique V, D or J segment in case of suboptimal sequencing accuracy. Nonetheless, we here show that targeted nanopore sequencing provides sufficient data to confidently map reads to unique genomic stretches within the Ig and TCR loci and to identify the underlying rearranged V, D or J segments by manual analysis of the provided genomic breakpoint coordinates. Key for the design of successful Ig/TCR MRD assays is the ability to reliably report the unique sequence of the newly formed breakpoint junction to facilitate the design of breakpoint-specific primers[396]. Furthermore, based on the speed of the assay, additional Sanger sequencing could be performed, if necessary, to obtain the exact sequence composition. Finally, blasting of the generated output against VDJ repertoire databases such as IGMT blast[416] could offer streamlined solutions for automated analyses of the underlying V(D)J configurations.

To implement this assay as a diagnostic tool further validations need to be performed. Thus far, we have retrospectively investigated ten samples. To accurately define the sensitivity of our assay, more samples need to be tested and additional MRD targets com-

pared. Furthermore, the implementation of additional recurrent fusion gene targets, such as *BCR-ABL1* or *RUNX1*, would broaden the target population of the assay[417]. A prospective side-by-side comparison of current diagnostic methods and our technique is required to confirm the advantages of our assay with respect to time, accuracy and sensitivity to identify appropriate MRD targets for PCR-based assays.
Taken together, we have developed a targeted nanopore sequencing assay for lymphoid leukemia patients to simultaneously identify recurrent fusion genes as well as the rearranged Ig and TCR repertoire. Our assay detects known and novel patient specific genomic MRD targets within 48 hours after sequencing library initiation, a time-frame comparable diagnostic methods are unable to meet.

MATERIAL AND METHODS

PATIENT MATERIAL

This study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice, and informed consent was obtained from all patients or their guardians. Collection and use of patient material was approved by the institutional review boards of the Princess Maxima Center for Pediatric Oncology in Utrecht, the Netherlands.

DNA-ISOLATIONS

Genomic DNA was extracted from blood or bone marrow with the QIAamp blood mini kit. Sample quality control was performed using a 4200 TapeStation System (Agilent), and DNA content was measured with a Qubit 3.0 Fluorometer (Thermo Fisher).

CRRNA DESIGN

crRNAs were designed as previously described[403]. In brief, the known target fusion partners were designated as a 5' or 3' fusion partner, dependent upon known literature. Furthermore, the most common breakpoint locations for *KMT2A*, *SIL*, and *TAL1* genes were extracted from a literature search and the most distal breakpoint locations were noted as extreme borders of the targeted area. For the Ig and TCR loci, crRNAs were designed adjacent to J segments, directing sequencing towards the (D)V segments.Therefore, genomic coordinates from the first to the last J segment were defined. If the unknown fusion partner or the V or D segments was the 5' partner, crRNAs were designed as the sequence present on the minus strand of the gene (5'->3') until the PAM sequence. If the unknown fusion partner or the V or D segments was the 3' partner, crRNAs were designed as the sequence present on the plus strand of the gene (5'->3') until the PAM sequence. Custom Alt-R® crRNAs were chosen with maximum on-target and off-target scores (IDT).

CAS9-ENRICHMENT AND NANOPORE SEQUENCING

Cas9 enrichment and Nanopore sequencing was performed as previously described[403]. In brief, approximately 1 µg of genomic DNA was dephosphorylated with Quick calf intestinal phosphatase (NEB) and CutSmart Buffer (NEB) for 10 minutes at 37 °C and inactivated for 2 minutes at 80 °C. crRNAs were resuspended in TE pH7.5 to 100 µM. For simultaneous targeting of multiple loci, crRNAs were pooled equimolarly to 100 µM. Ribonucleoprotein complexes (RNPs) were prepared by mixing 100 µM equimolarialy pooled crRNA pools with 100 µM tracrRNA (IDT) and duplex buffer (IDT), incubated for 5 minutes at 95°C and thereafter cooled to room temperature. 10 µM RNPs were mixed with 62 µM HiFiCas9 (IDT) and 1x CutSMart buffer (NEB) and incubated at RT for 15 minutes to produce Cas9 RNPs. Dephosphorylated DNA samples and Cas9 RNPs were mixed with 10mM dATP and Taq polymerase (NEB) at 37 °C for 15 minutes and 72 °C for 5 minutes to facilitate cutting of the genomic DNA and dA-tailing. Adaptor ligation mix was prepared by mixing Ligation Buffer (SQK-LSK109, ONT), Next Quick T4 DNA Ligase (NEB) and Adaptor Mix (SQK-LSK109, ONT). The mix was carefully applied to the processed DNA sample without vortexing and incubated at room temperature for 25 minutes. DNA was washed and bound to beads by adding TE pH8.0 and 0.3 x volume AMPure XP beads (Agencourt) and incubated for 10 minutes at room temperature. Fragments below 3 kb were washed away by washing the bead-bound solution twice with Long Fragment Buffer (SQK-LSK109, ONT). Enriched library was released from the beads with Elution Buffer (SQK-LSK109, ONT). Enriched library concentration was measured with the Qubit Fluorometer 3.0 (Thermo Fisher).
The library from one tumour sample was loaded onto one Flow Cell (R 9.4, ONT) according to the manufacturer's protocol. Sequencing was performed on a GridION X5 instrument (ONT) and basecalling was performed by Guppy (ONT).

DATA ANALYSIS

Nanopore sequencing data were mapped against the reference genome (GRCh38) with minimap2 (v2.12[344]) with parameters '*-x map-ont -a --MD*'. Fusion genes were detected and reported as previously described with NanoFG[403]. Mapped BAM files were split in the different IG and TR loci with samtools (v1.9[418]). SV calling per loci was performed with Sniffles (v1.0.12[342]) with parameters '*-n -1 --report_BND --genotype -s 2 --max_num_splits 100 -d 100 -l 20*'. Primer design for SV validation is automatized by using Primer3 (v1.1.4[349]) with a product size range of 30-230 bp. Additionally, for each SV detected, supporting reads were used for consensus calling with 4 rounds of read overlap with minimap2 (v2.12[344]) and Racon (v. 1.4.15[419]) followed by polishing with Medaka (v.1.0.3, https://github.com/nanoporetech/medaka).

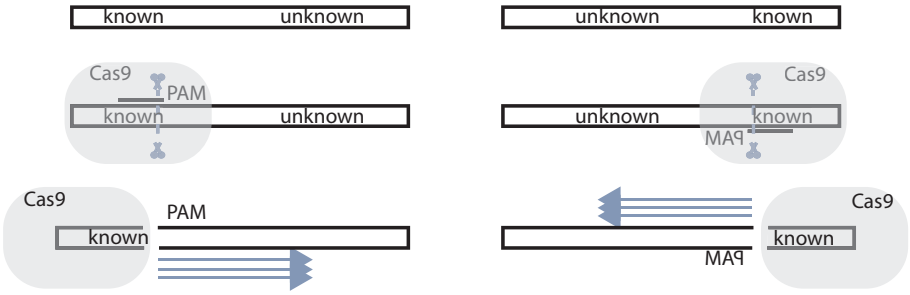
Average coverages were calculated with Sambamba (v0.6.5[345]) with parameters *'depth base --min-coverage=0'*. On target coverage was defined as the average coverage at cut-site plus 20-40bps in the desired sequencing direction. Ig and TCR loci coordinates were obtained from Gencode v.29[4].

PCR VALIDATION

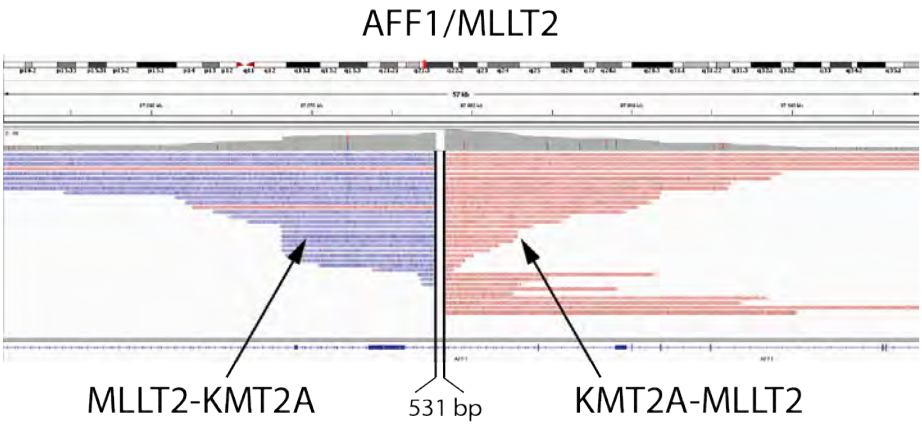
PCR validations were performed using the AmpliTaq Gold DNA Polymerase (Thermofisher) protocol utilizing 50 ng DNA. The PCR conditions were as follows: 95 °C for 5 minutes, followed by 14 cycles of 95 °C for 30 seconds, 65 - 58 °C decreasing n 0.5 °C increments, and 72 °C for one minute, followed by a further 24 cycles of 95 °C for 30 seconds, 58 °C for 30 seconds, and 72 °C for 1 minute, and completed by 72 °C for 10 minutes a hold at 10 °C.
PCR products were then run on a 2% agarose gel at 100 V for one hour for assessment of assay specificity.

LIST OF SUPPLEMENTARY DATA

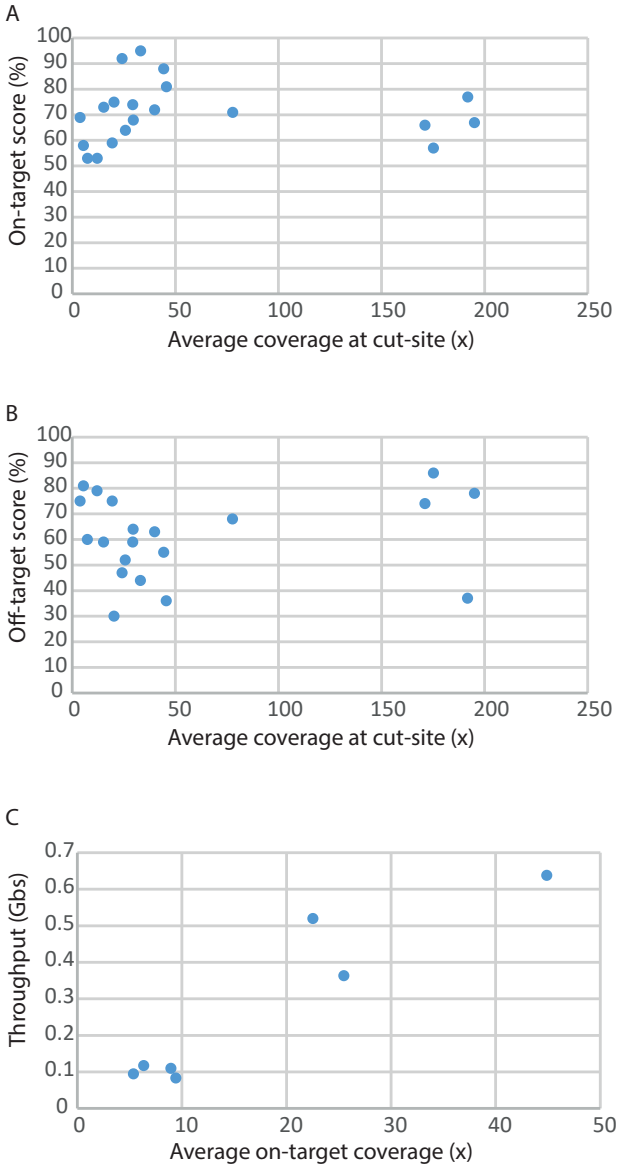
**Table S1-3 are available online at: https://tinyurl.com/Ch6Suppl or scanning the QR code below*



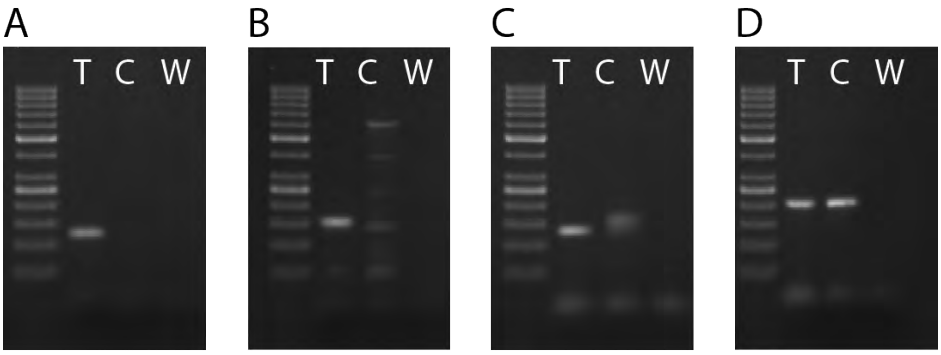
Supplementary Figure 1: Schematic overview of targeted and directional sequencing. **(A)** Known sequence adjacent to the unknown sequence of interest is necessary for assay design. **(B)** Cas9 (grey box) is targeted through crRNAs (line) and the PAM sequence to the known sequence and introduces a double-strand break (scissors) in the DNA. **(C)** Cas9 blocks the PAM-distal side and sequencing reads (arrows) are directed towards the unknown sequence.



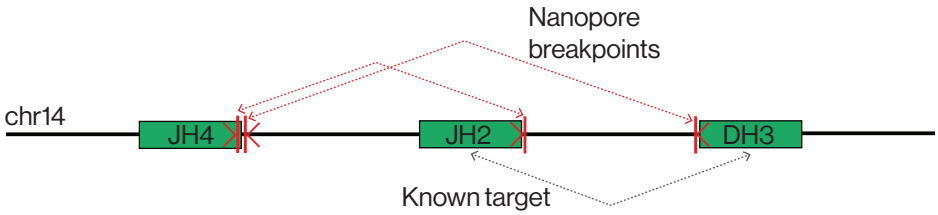
Supplementary Figure 2: Reciprocal KMT2A fusion reads within the MLLT2 locus (B-ALL5). IGV screenshot of the B-ALL5 sample of fusion-spanning reads supporting the MLLT2-KMT2A fusion **(blue)** and KMT2A-MLLT2 fusion **(red)** within the MLLT2 locus. Note that reads start 531 bp apart.



Supplementary Figure 3: On-target coverage depends on total throughput. Relation between average coverage at cut-site and (**A**) on-target and (**B**) off-target scores of the crRNAs show no obvious correlation. (**C**) We observe a correlation between average on-target coverage and total sequencing throughput.



Supplementary Figure 4: Examples of breakpoint-spanning amplification bands. PCR with breakpoint-spanning primers is performed on tumor sample (T), polyclonal healthy donor DNA (C) and water (W). (**A**-**C**) Successful PCR assay since amplification band is only present in tumor sample (**A**), presents with a smear in control sample (**B**), or shows a difference in size between tumor and control sample (**C**). (**D**) Amplification band in tumor as well as control sample, indicating no specificity of the rearrangement to the patient sample.



Supplementary Figure 5: Alternative configuration of DH3-JH2 target for sample B-ALL2. Despite the already known MRD target Dh3-Jh2 in sample B-ALL5, nanopore sequencing data links the Dh3 and Jh2 loci through two chained rearrangements involving also the Jh4 locus.

AUTHOR'S CONTRIBUTIONS

CS and GM conceived the study. CS designed the assays and IR, KD and CS performed the experiments. LW provided samples, clinical data and input about diagnostic relevance. JEV-I performed the bioinformatic analysis. CS and JEV-I analyzed the data. CS, JEV-I, GH and GM interpreted the data. CS and JEV-I wrote the manuscript which was edited by LW, BT, EV, GvH and GM.

ACKNOWLEDGMENTS

We thank all patients for providing the clinical specimens to perform this study and the Princess Maxima Center for providing these samples. We thank the van Haaften group for discussions and support. We thank the Utrecht Sequencing Facility for providing the nanopore sequencing.

7

GENERAL DISCUSSION

Cancer is a genomic disease and tumors acquire somatic mutations throughout their origin and evolution. Some of these mutations give a tumor growth advantage over normal cells (driver mutations) and most do not have selective effect (passenger mutations). Regardless, somatic mutations, including structural variants (SVs), uniquely identify and characterize a tumor and can be exploited in different ways. Mainly through targeted treatments that inhibit the growth advantages conferred by driver mutations, killing specifically the cancer cells. However, passenger events can also be used as cancer tags to differentiate tumors from normal cells and measure cancer dynamics. It is therefore possible to turn cancer's biggest strengths, the acquired mutations and growth advantages, into its pitfall by leveraging genomic technologies for research and patients' benefit. In this final chapter, I summarize several technological contributions presented in the previous chapters and discuss their potential and limitations.

ORGANOIDS AS A TOOL FOR OVARIAN CANCER RESEARCH

Patient-derived cell lines are useful models that are nowadays relatively easy to establish, maintain, genetically manipulate and are amenable to large high-throughput drug screening. However, more complex models represent the complexity and heterogeneity of tumors better. Patient-derived xenografts are *in vivo* mice models that provide a more realistic tumor environment, therefore increasing the predictive value of drug screens. They are however expensive, labor-intense and less suitable for large screening. Moreover, these xenografts might evolve towards a mouse-specific setting that, despite being *in vivo* models, deviates from the original tumor's biology[115]. Patient-derived organoids (PDOs) are faithful *in vitro* representations of the tumor characteristics, such as tumor heterogeneity and genomic traits, while allowing for high-throughput drug screening and genetic manipulation. PDOs have the potential to make personalized and targeted treatments for patients with ovarian cancer (OC) a reality and to be important model systems to improve our knowledge of cancer biology.

In **chapter 2** we describe a biobank of OC PDOs. This biobank represents multiple different OC subtypes and is largely characterized at the histological, genomic, epigenomic and transcriptomic level, demonstrating the faithfulness to the original tumor material and stability over time. Furthermore, creating PDO biobanks of OC and other cancer types provides platforms for large scale drug screening and genomic and functional studies, as previously shown for colorectal cancer. The establishment of healthy fallopian tube organoids and patient-derived OC organoids has been reported independently by other groups[279,309,310,420]. All these studies show similar culture success rates to ours, which are lower when the obtained resected tumor material has been previously exposed to treatment, as is the case for most of our PDO lines. These independent studies also show

that the established PDOs are representative of the original tumor, with shallow genetic characterizations using targeted gene panels or whole-exome sequencing. We, however, provide a more comprehensive characterization with the use of whole genome sequencing (WGS) for genomic, RNA-seq for transcriptomic and even methylation arrays for epigenomic characterization, showing that the PDOs are representative of the corresponding tumor and that the heterogeneity that characterizes OC is captured in the biobank. Moreover, by characterizing PDOs after long term passaging we show that those features are maintained well over time. We also show examples of potential applications of the biobank by using genome editing techniques to model the origin of OC and also performing drug screening with drugs that are commonly used in OC treatment, even assessing their response *in vivo* after xenotransplantation into mice.

Another strength of our biobank is the establishment of PDO lines from different tumor locations of the same patient. While in **chapter 2** we describe how the intra-patient genomic heterogeneity is preserved, in **chapter 3** we expand to study how these related locations can respond differently to the same drug. There, we found differential responses to at least one drug in each of these patients. This suggests that tumor heterogeneity can also involve multiple locations of the same patient, influencing treatment efficacy and resistance development. Establishing organoids from all the tumor locations from a patient can enable the screening of multiple drugs and the selection of different effective treatments that work for all of them. In order to achieve this stage, fast PDO culture establishment with high success rate to ensure that all locations from all patients can be represented in a clinically relevant timeframe, as well as standardized drug assays and well-defined controls to evaluate organoid response are needed.

There are several possibilities to improve success rates when establishing organoid lines from patients with OC. Most of the tumors we sampled for organoid derivation had been previously exposed to chemotherapy, since the tissue samples were obtained during interval debulking as part of the standard treatment for OC. Therefore, the tumor viability might be already compromised prior to culture establishment due to the genomic damage induced by the chemotherapeutic agents. To overcome this, obtaining biopsies prior to chemotherapy exposure or selection of areas with higher tumor purity and viability prior to plating, would be ideal. Furthermore, different growth factors used in the medium might be driving selection for certain samples depending on their specific molecular characteristics. For example, nicotinamide is a growth factor used in our organoid medium and also in previous organoid biobanks. However, nicotinamide is also a PARP-inhibitor, which is a compound that inhibits DNA repair by the homologous recombination (HR) machinery. PARP-inhibitors are thus effective drugs for HR-deficient tumors, such as those with BRCA1 or BRCA2 inactivation. Nevertheless, in this case, it might be

inhibiting the growth of HR-deficient organoids that are then under-represented in our biobank. None of the PDOs in our biobank was defined as HR-deficient according to CHORD, an algorithm that detects genomic evidence of HR-deficiency[421]. Optimization of the culture media and further investigation of possible selection effects might then improve culture success rates.

A different approach than biobanking for the use of PDOs is the establishment of short-term cultures that can be used to rapidly screen relevant drugs for a particular patient. In **chapter 3** we perform this rapid organoid establishment and drug screening for one sample, obtaining response results within three weeks after tissue collection. Additionally, other efforts show that it is possible to perform rapid and short-term culturing and drug screening of OC PDOs or PDO-like cultures, typically within weeks after tissue collection, albeit sacrificing extensive molecular characterization[279,309,310,420]. Nevertheless, larger cohorts of OC patients are needed to demonstrate that PDO models are representative and predictive of the drug responses and clinical outcomes in patients, as proven in colorectal cancer[138,281]. Therefore, to maximize the clinical and research utility of PDOs rapid short term cultures for relevant drug screening should be established parallel to the expansion for biobanking.

Finally, PDOs overcome practical challenges of cancer research using real tumor samples, such as sufficient material available and possibility of genomic modification and engineering, while providing a more faithful model than traditional cell lines as discussed above. It is essential to capture all the cancer complexity in OC and other cancer types to advance the use of PDOs as cancer research models. For example, understanding the origin of chemotherapy resistance and its prediction or detection is of utter importance in OC, since most patients respond well to initial chemotherapy treatments but relapse with resistant tumors. Paired organoids from patients before and after chemotherapy treatment would be relevant to study the mechanisms behind the acquisition of resistance. Another limitation of PDOs is the lack of a functional immune system, which is essential to understand therapy response and tumor evolution[422]. To overcome this limitation and bring the PDO *in vitro* system to the *in vivo* situation, co-culture systems of PDOs and immune cells are being developed[423,424]. Furthermore, in order to better understand the culture effects and differential drug responses in these organoid cultures, it would be interesting to further characterize these PDOs with multi-omics approaches that also detect presence of proteins, metabolites or neo-antigens. Perhaps a more extensive characterization beyond genomics might find explanations for the differential responses that we described in **chapter 3**.

CHALLENGES FOR SOMATIC STRUCTURAL VARIATION DETECTION

The role of somatic SVs in cancer has remained more concealed than for other mutation types. The lack of gold-standard benchmarks and best-practices workflows to detect and filter SV calls have reduced the impact of technological advances in the cancer SV field. Due to the chromosomal instability of tumors they rapidly accumulate more SVs than normal cells, often in clustered events. Therefore, care should be taken when transitioning from germline SV detection towards somatic SV detection. It is essential to perform joint calling of tumor and normal pairs to discard SV events present in the germline. Additionally, somatic SV calls might present different characteristics than germline resulting from specific mechanisms of DNA damage active in tumorigenesis. Lastly, for somatic SV discovery a "kitchen sink" approach is often applied: try every available algorithm at the data and use the ensemble result. Apart from the computational concerns of this approach, Cameron *et al.*[359] demonstrated that this ensemble method does not necessarily improve sensitivity and specificity of somatic detection. There is therefore a need for enhancement of methods in the field of somatic SV detection, and truth sets and benchmarks are essential for these developments.

Gold reference truth sets of SVs have been established in germline individuals using integrative approaches with orthogonal genomic data from multiple sequencing technologies. Chaisson *et al.*[199] and Zook *et al.*[213] show the importance of using long-read data for comprehensive germline SV detection, especially when SVs fall in tandem repeat regions and insertions. Both studies, however, emphasize the impracticality of performing this type of orthogonal approach in routine clinical practice, given the elevated costs and computational difficulties for data integration. Nevertheless, these benchmarks are essential to compare methods, algorithms and technologies and to establish optimal cost and sensitivity thresholds for SV detection. In **chapter 4**, we followed a similar orthogonal approach to establish a truth set for somatic SV calling using the COLO829 melanoma and paired COLO829BL normal cell lines derived from the same individual. Previous efforts have focused on SNVs and have therefore used only short-read data. Arora *et al.*[214] used the COLO829 cell lines and two others to benchmark two short-read sequencing platforms. Other somatic SV benchmarking efforts have made use of simulated data[330,331,359] or mouse genomic data[332], while stressing the need for a high quality truth set to enhance somatic SV method development.

Surprisingly and in contrast to germline SV characterization studies, the contribution of long-reads in our somatic SV truth set was relatively low and the truth set was almost completely resolved with the short-read dataset alone. There are two possible reasons for this. First, somatic SVs might have fundamental differences compared to germline SVs.

Due to the increased DNA damage or reduced DNA repair mechanisms in cancer cells, somatic SVs accumulate in more regions than just repeat-rich regions, perhaps mitigating the added value of long-reads in these regions. Second, algorithms for somatic SV detection are more mature for short-read data than for long-read data. For example, a joint caller that identifies somatic SVs using simultaneous input data from tumor and matched normal from nanopore sequencing data does not exist, hindering the somatic SV discovery process for this type of data. Previous studies on somatic SV detection with nanopore data have reported high numbers of somatic SVs detected only by long-reads[208,209]. However, in contrast to our study, no experimental validation of the novel SV calls was performed in these studies. It is therefore likely that a large fraction of these are false positives. Nevertheless, with more optimal algorithms and experimental methods emerging, long-read technologies will likely improve their capability of somatic SV detection, and with our truth set we provide a resource to measure and benchmark this.

APPLICATIONS OF SOMATIC SV AND NANOPORE SEQUENCING IN CANCER

Despite the current limitations of long-reads for somatic structural variation detection described above, long-reads in general and nanopore sequencing in particular still can provide novel contributions to cancer research and medicine. For example, application of long-read sequencing enables the identification of germline cancer-predisposing SVs that were not elucidated with short-read sequencing[425]. Also, nanopore sequencing is currently being used to close current gaps in the human reference genome, particularly regarding centromeres. An international consortium has managed to provide telomere-to-telomere chromosome assemblies of chromosomes 8[426] and X[427], and has ambitions to completely sequence end-to-end a human genome. Furthermore, another study by Ebbert *et al.*[178] reveals that with the use of linked- and long-read technologies a larger fraction of the genome becomes available, revealing disease-relevant genes that remained inaccessible by other technologies. More complete genome references and longer reads, combined with optimal algorithms, might also aid to complete the somatic SV truth set that we present in **chapter 4**, where several copy number alterations that involve centromeres still miss an associated SV breakpoint.

We leveraged other characteristics of nanopore sequencing apart from the long-read lengths: the simplicity and the rapid nature of the sequencing approach. Oxford Nanopore Technologies markets sequencing devices of small size, ranging between a USB stick to a small desktop machine. They attempt to bring sequencing closer to the patient and promote in-house sequencing opposed to large centralized facilities, reducing the logistics needed in diagnostic laboratories. In **chapters 5** and **6** we developed two assays to profile somatic SVs from tumor-only nanopore sequencing data. We show that the somatic SVs

that we detect can be utilized as biomarkers to monitor tumor dynamics and to detect minimal residual disease. The use of somatic SVs as tumor-specific biomarkers is not a novel approach and has been described before[69,165,166]. The main advantage of somatic SVs as opposed to other types of genomic biomarkers is that the specific break-junction created can be detected by high-sensitivity PCR-based methods. In the studies referred, tumor-specific somatic SVs are detected approximately one month after tissue collection. We, however, provide the tumor-specific biomarkers within a few days after tissue collection, given rapid nanopore library preparation and sequencing and the immediate availability of the resulting data. This is particularly relevant for evaluation of treatment response in patients by measuring the reduction in circulating tumor DNA (ctDNA).

To extract relevant somatic SVs from noisy nanopore sequencing data from low coverage tumor samples without a germline control we used two different approaches. First, in **chapter 5**, we used random forest classification to distinguish real SVs from false positives, followed by ranking of the positive candidates by likelihood of being somatic, based on SV size and labelling of known germline variants. We are therefore not distinguishing between driver or passenger somatic SVs. Second, in **chapter 6**, we devise a wet-lab protocol with CRISPR-Cas9 technology to enrich the sequencing of regions that contain the relevant somatic SVs used for MRD tracing in leukemia, such as known fusion genes and the Ig and TCR loci. We are therefore targeting known leukemia SV driver events in this case.

In both of these assays, a main drawback is that they cannot be used for comprehensive genome-wide SV characterization, since we perform either low-coverage genome sketching or targeted sequencing. Also, distinguishing true somatic SVs is a significant issue regardless of the cancer driver status of the event, as these SVs might be lost during tumor evolution especially under the strong pressure selection amid treatment. We overcome this risk by selecting multiple somatic SVs in each sample, therefore decreasing the change of false negatives at the tracking step - when a biomarker is not detected even though there is tumor presence. We could find at least 5 biomarkers in each sample in **chapter 5**, which is in line with previous studies[69,165,166]. However, more data and larger cohorts are needed to determine an optimal number of SVs to monitor per patient, as well as to establish the sensitivity, specificity and prognostic value of the assays.

Further potential improvements of our assays include the creation of large panels of germline and somatic SVs based on population approaches. With such panels, we would be able to better distinguish germline from somatic SVs from our tumor-only low-depth sequencing approaches. For example, a recent project by the Genome Aggregation Database (gnomAD) collected 433,371 germline SVs from 14,891 genome[428]. Importantly,

these resources need to be public and fully open access for the benefit of technological advance. Also, other groups have developed computational enrichment approaches with nanopore sequencing. Rather than using wet-lab Cas9-based enrichment, specific targeted regions are sequenced by mapping the read in real-time while it traverses the pore and rejecting it when it is not part of the target region[429,430]. This dry enrichment technique was used by Miller *et al.*[431] to enrich suspected regions and discover pathogenic SVs in Mendelian disorders. By combining these computational enrichment with our approaches, perhaps more target regions and higher enrichments in cancer-relevant regions could be achieved, increasing confidence in the biomarkers that we detect and reducing post-analysis validation.

FINAL REMARKS

In this thesis I introduced and discussed several approaches that leverage state of the art genomic technology for the advancement of cancer research and diagnostics. I expect that the assays and resources presented here will be of use to the international cancer genomics and cancer research community. Our OC PDO biobank serves the OC research community and shows potential to also impact the OC clinical care. Continued establishment and expansion of PDOs will help reach this potential, as discussed above. Furthermore, I leverage SVs, which have been traditionally overlooked in cancer genomics, and long read sequencing technologies and provide community resources and personalized medicine approaches that use somatic SVs, overcoming previous technological limitations. With the speed that genomic technology advances, maybe these approaches will become outdated in the next few years with new technologies, methods and algorithms. Nevertheless, the contributions presented here are relevant to drive this technological advance by pushing the limits of the state of the art of cancer genomics.

ADDENDUM

REFERENCES

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738 (1953).
2. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. Nature 171, 738–740 (1953).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
4. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
6. Levy, S. *et al.* The diploid genome sequence of an individual human. PLoS Biol. 5, e254 (2007).
7. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58 (2010).
8. Eichler, E. E. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N. Engl. J. Med. 381, 64–74 (2019).
9. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674 (2011).
10. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
11. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. 14, 125–138 (2013).
12. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376 (2011).
13. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).
14. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. Science 349, (2015).
15. Chiang, C. *et al.* The impact of structural variation on human gene expression. Nat. Genet. 49, 692–699 (2017).
16. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712 (2010).
17. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11, R52 (2010).
18. Cancer today. http://gco.iarc.fr/today/home.
19. Consortium, T. I. P.-C. A. of W. G. & The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020).

20. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. Nat. Rev. Cancer 20, 555–572 (2020).
21. Rahman, N. Realizing the promise of cancer predisposition genes. Nature 505, 302–308 (2014).
22. Lane, D. P. p53, Guardian of the genome. Nature 358, 15–16 (1992).
23. Holderfield, M., Deuker, M. M., McCormick, F. & McMahon, M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nat. Rev. Cancer 14, 455–467 (2014).
24. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005).
25. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 18, 1051–1063 (2008).
26. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).
27. International Cancer Genome Consortium *et al.* International network of cancer genome projects. Nature 464, 993–998 (2010).
28. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).
29. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575, 210–216 (2019).
30. Gröbner, S. N. *et al.* Author Correction: The landscape of genomic alterations across childhood cancers. Nature 559, E10 (2018).
31. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature 555, 371–376 (2018).
32. Wong, M. *et al.* Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. Nat. Med. (2020) doi:10.1038/s41591-020-1072-4.
33. Vis, D. J. *et al.* Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. Ann. Oncol. 28, 1145–1151 (2017).
34. Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing Genome-Driven Oncology. Cell 168, 584–599 (2017).
35. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. Nat. Rev. Clin. Oncol. 15, 353–365 (2018).
36. Nangalia, J. & Campbell, P. J. Genome Sequencing during a Patient's Journey through Cancer. New England Journal of Medicine 381, 2145–2156 (2019).
37. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N. Engl. J. Med. 344, 1031–1037 (2001).
38. Shaw, A. T. *et al.* Crizotinib in ROS1-rearranged non-small-cell lung cancer. N. Engl. J. Med. 371, 1963–1971 (2014).

39. Robert, C. *et al.* Improved overall survival in melanoma with combined dabrafenib and trametinib. N. Engl. J. Med. 372, 30–39 (2015).
40. De Roock, W. *et al.* Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncol. 11, 753–762 (2010).
41. Kumar-Sinha, C. & Chinnaiyan, A. M. Precision oncology in the age of integrative genomics. Nat. Biotechnol. 36, 46–60 (2018).
42. Burd, A. *et al.* Precision medicine treatment in acute myeloid leukemia using prospective genomic profiling: feasibility and preliminary efficacy of the Beat AML Master Trial. Nat. Med. (2020) doi:10.1038/s41591-020-1089-8.
43. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).
44. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 350, 207–211 (2015).
45. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science 357, 409–413 (2017).
46. Robson, M., Goessl, C. & Domchek, S. Olaparib for Metastatic Germline BRCA-Mutated Breast Cancer. The New England journal of medicine vol. 377 1792–1793 (2017).
47. van der Velden, D. L. *et al.* The Drug Rediscovery protocol facilitates the expanded use of existing anticancer drugs. Nature 574, 127–131 (2019).
48. Garraway, L. A. & Jänne, P. A. Circumventing cancer drug resistance in the era of personalized medicine. Cancer Discov. 2, 214–226 (2012).
49. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 547, 217–221 (2017).
50. Fritsch, E. F., Hacohen, N. & Wu, C. J. Personal neoantigen cancer vaccines: The momentum builds. Oncoimmunology 3, e29311 (2014).
51. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat. Rev. Genet. 20, 71–88 (2019).
52. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat. Rev. Cancer 17, 223–238 (2017).
53. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 359, 926–930 (2018).
54. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. Nat. Med. 14, 985–990 (2008).
55. Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci. Transl. Med. 4, 162ra154 (2012).
56. Mattos-Arruda, L. D. *et al.* Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. Annals of Oncology 25, 1729–1735 (2014).

57. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature 545, 446–451 (2017).
58. Parkinson, C. A. *et al.* Exploratory Analysis of TP53 Mutations in Circulating Tumour DNA as Biomarkers of Treatment Response for Patients with Relapsed High-Grade Serous Ovarian Carcinoma: A Retrospective Study. PLoS Med. 13, e1002198 (2016).
59. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci. Transl. Med. 6, 224ra24 (2014).
60. Leighl, N. B. *et al.* Molecular testing for selection of patients with lung cancer for epidermal growth factor receptor and anaplastic lymphoma kinase tyrosine kinase inhibitors: American Society of Clinical Oncology endorsement of the College of American Pathologists/ International Association for the study of lung cancer/association for molecular pathology guideline. J. Clin. Oncol. 32, 3673–3679 (2014).
61. Warren, J. D. *et al.* Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. BMC Med. 9, 133 (2011).
62. Gormally, E. *et al.* TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. Cancer Res. 66, 6871–6876 (2006).
63. Beaver, J. A. *et al.* Detection of cancer DNA in plasma of patients with early-stage breast cancer. Clin. Cancer Res. 20, 2643–2650 (2014).
64. Dawson, S.-J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. N. Engl. J. Med. 368, 1199–1209 (2013).
65. Gray, E. S. *et al.* Circulating tumor DNA to monitor treatment response and detect acquired resistance in patients with metastatic melanoma. Oncotarget 6, 42008–42018 (2015).
66. Tie, J. *et al.* Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci. Transl. Med. 8, 346ra92 (2016).
67. Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci. Transl. Med. 7, 302ra133 (2015).
68. Reinert, T. *et al.* Analysis of circulating tumour DNA to monitor disease burden following colorectal cancer surgery. Gut 65, 625–634 (2016).
69. Olsson, E. *et al.* Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. EMBO Mol. Med. 7, 1034–1047 (2015).
70. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886 (2015).
71. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. Science 362, 911–917 (2018).
72. Vaughan, S. *et al.* Rethinking ovarian cancer: recommendations for improving outcomes. Nat. Rev. Cancer 11, 719–725 (2011).
73. Bowtell, D. D. *et al.* Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nat. Rev. Cancer 15, 668–679 (2015).
74. Karnezis, A. N., Cho, K. R., Gilks, C. B., Pearce, C. L. & Huntsman, D. G. The disparate ori-

gins of ovarian cancers: pathogenesis and prevention strategies. Nat. Rev. Cancer 17, 65–74 (2017).
75. Ng, A. & Barker, N. Ovary and fimbrial stem cells: biology, niche and cancer origins. Nat. Rev. Mol. Cell Biol. 16, 625–638 (2015).
76. Torre, L. A. *et al.* Ovarian cancer statistics, 2018. CA: A Cancer Journal for Clinicians 68, 284–296 (2018).
77. Network, T. C. G. A. R. & The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615 (2011).
78. Patch, A.-M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. Nature 521, 489–494 (2015).
79. Prat, J., Ribé, A. & Gallardo, A. Hereditary ovarian cancer. Human Pathology 36, 861–870 (2005).
80. Walsh, T. *et al.* Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. 108, 18032–18037 (2011).
81. Toss, A. *et al.* Hereditary ovarian cancer: not only BRCA 1 and 2 genes. Biomed Res. Int. 2015, 341723 (2015).
82. Wang, Y. K. *et al.* Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. Nat. Genet. 49, 856–865 (2017).
83. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434, 917–921 (2005).
84. Pommier, Y., O'Connor, M. J. & de Bono, J. Laying a trap to kill cancer cells: PARP inhibitors and their mechanisms of action. Sci. Transl. Med. 8, 362ps17 (2016).
85. Mateo, J. *et al.* A decade of clinical development of PARP inhibitors in perspective. Ann. Oncol. 30, 1437–1447 (2019).
86. Hoogstraat, M. *et al.* Genomic and transcriptomic plasticity in treatment-naive ovarian cancer. Genome Research 24, 200–211 (2014).
87. Caponigro, G. & Sellers, W. R. Advances in the preclinical testing of cancer therapeutic hypotheses. Nat. Rev. Drug Discov. 10, 179–187 (2011).
88. Voskoglou-Nomikos, T., Pater, J. L. & Seymour, L. Clinical predictive value of the *in vitro* cell line, human xenograft, and mouse allograft preclinical cancer models. Clin. Cancer Res. 9, 4227–4239 (2003).
89. Pauli, C. *et al.* Personalized and Cancer Models to Guide Precision Medicine. Cancer Discov. 7, 462–477 (2017).
90. Gey, G. O., Coffman, W. D. & Kubicek, M. T. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res. 12, 264–265 (1952).
91. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature 500, 207–211 (2013).
92. Scherer, W. F., Syverton, J. T. & Gey, G. O. Studies on the propagation *in vitro* of poliomyelitis

viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. J. Exp. Med. 97, 695–710 (1953).
93. Beskow, L. M. Lessons from HeLa Cells: The Ethics and Policy of Biospecimens. Annu. Rev. Genomics Hum. Genet. 17, 395–417 (2016).
94. Masters, J. R. HeLa cells 50 years on: the good, the bad and the ugly. Nat. Rev. Cancer 2, 315–319 (2002).
95. Skloot, R. The Immortal Life of Henrietta Lacks. (Pan Publishing, 2011).
96. Kamb, A. What's wrong with our cancer models? Nat. Rev. Drug Discov. 4, 161–165 (2005).
97. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anti-cancer drug sensitivity. Nature 483, 603–607 (2012).
98. Verschraegen, C. F. *et al.* Establishment and characterization of cancer cell cultures and xenografts derived from primary or metastatic Mullerian cancers. Clin. Cancer Res. 9, 845–852 (2003).
99. Ince, T. A. *et al.* Characterization of twenty-five ovarian tumour cell lines that phenocopy primary tumours. Nat. Commun. 6, 7419 (2015).
100. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. Nat. Commun. 4, 2126 (2013).
101. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. Nature 560, 325–330 (2018).
102. Tentler, J. J. *et al.* Patient-derived tumour xenografts as models for oncology drug development. Nat. Rev. Clin. Oncol. 9, 338–350 (2012).
103. Frese, K. K. & Tuveson, D. A. Maximizing mouse cancer models. Nat. Rev. Cancer 7, 654–658 (2007).
104. Siolas, D. & Hannon, G. J. Patient-derived tumor xenografts: transforming clinical samples into mouse models. Cancer Res. 73, 5315–5319 (2013).
105. Aparicio, S., Hidalgo, M. & Kung, A. L. Examining the utility of patient-derived xenograft mouse models. Nat. Rev. Cancer 15, 311–316 (2015).
106. Heo, E. J. *et al.* Patient-Derived Xenograft Models of Epithelial Ovarian Cancer for Preclinical Studies. Cancer Res. Treat. 49, 915–926 (2017).
107. Liu, J. F. *et al.* Establishment of Patient-Derived Tumor Xenograft Models of Epithelial Ovarian Cancer for Preclinical Evaluation of Novel Therapeutics. Clin. Cancer Res. 23, 1263–1273 (2017).
108. Bankert, R. B. *et al.* Humanized mouse model of ovarian cancer recapitulates patient solid tumor progression, ascites formation, and metastasis. PLoS One 6, e24420 (2011).
109. Ricci, F. *et al.* Patient-derived ovarian tumor xenografts recapitulate human clinicopathology and genetic alterations. Cancer Res. 74, 6980–6990 (2014).
110. Topp, M. D. *et al.* Molecular correlates of platinum response in human high-grade serous ovarian cancer patient-derived xenografts. Mol. Oncol. 8, 656–668 (2014).
111. George, E. *et al.* A patient-derived-xenograft platform to study BRCA-deficient ovarian can-

cers. JCI Insight 2, e89760 (2017).
112. Harris, F. R. *et al.* Targeting HER2 in patient-derived xenograft ovarian cancer models sensitizes tumors to chemotherapy. Mol. Oncol. 13, 132–152 (2019).
113. Guffanti, F. *et al.* Platinum sensitivity and DNA repair in a recently established panel of patient-derived ovarian carcinoma xenografts. Oncotarget 9, 24707–24717 (2018).
114. Hidalgo, M. *et al.* Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov. 4, 998–1013 (2014).
115. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution. Nat. Genet. 49, 1567–1575 (2017).
116. Woo, X. Y. *et al.* Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. bioRxiv (2020) doi:10.1101/861393.
117. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. Nature 459, 262–265 (2009).
118. Clevers, H. Modeling Development and Disease with Organoids. Cell 165, 1586–1597 (2016).
119. Kessler, M. *et al.* The Notch and Wnt pathways regulate stemness and differentiation in human fallopian tube organoids. Nat. Commun. 6, 8989 (2015).
120. Sato, T. *et al.* Long-term Expansion of Epithelial Organoids From Human Colon, Adenoma, Adenocarcinoma, and Barrett's Epithelium. Gastroenterology 141, 1762–1772 (2011).
121. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).
122. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science 358, 234–238 (2017).
123. Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. Nature 521, 43–47 (2015).
124. Weeber, F. *et al.* Preserved genetic diversity in organoids cultured from biopsies of human colorectal cancer metastases. Proc. Natl. Acad. Sci. U. S. A. 112, 13308–13311 (2015).
125. Roerink, S. F. *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. Nature 556, 457–462 (2018).
126. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. Cell 160, 324–338 (2015).
127. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. Cell 159, 176–187 (2014).
128. van de Wetering, M. *et al.* Prospective derivation of a living organoid biobank of colorectal cancer patients. Cell 161, 933–945 (2015).
129. Fujii, M. *et al.* A Colorectal Tumor Organoid Library Demonstrates Progressive Loss of Niche Factor Requirements during Tumorigenesis. Cell Stem Cell 18, 827–838 (2016).
130. Huang, L. *et al.* Ductal pancreatic cancer modeling and drug screening using human pluripotent stem cell- and patient-derived tumor organoids. Nat. Med. 21, 1364–1371 (2015).

131. Seino, T. *et al.* Human Pancreatic Tumor Organoids Reveal Loss of Stem Cell Niche Factor Dependence during Disease Progression. Cell Stem Cell 22, 454–467.e6 (2018).
132. Driehuis, E. *et al.* Pancreatic cancer organoids recapitulate disease and allow personalized drug screening. Proc. Natl. Acad. Sci. U. S. A. (2019) doi:10.1073/pnas.1911273116.
133. Broutier, L. *et al.* Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. Nat. Med. 23, 1424–1435 (2017).
134. Driehuis, E. *et al.* Oral Mucosal Organoids as a Potential Platform for Personalized Cancer Therapy. Cancer Discov. 9, 852–871 (2019).
135. Sachs, N. *et al.* A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. Cell 172, 373–386.e10 (2018).
136. Drost, J. & Clevers, H. Organoids in cancer research. Nat. Rev. Cancer 18, 407–418 (2018).
137. Neal, J. T. *et al.* Organoid Modeling of the Tumor Immune Microenvironment. Cell 175, 1972–1988.e16 (2018).
138. Vlachogiannis, G. *et al.* Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. Science 359, 920–926 (2018).
139. Ganesh, K. *et al.* A rectal cancer organoid platform to study individual responses to chemoradiation. Nat. Med. 25, 1607–1614 (2019).
140. Yao, Y. *et al.* Patient-Derived Organoids Predict Chemoradiation Responses of Locally Advanced Rectal Cancer. Cell Stem Cell 26, 17–26.e6 (2020).
141. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. Nat. Rev. Clin. Oncol. 15, 139–150 (2018).
142. Cheng, C. *et al.* Whole-Genome Sequencing Reveals Diverse Models of Structural Variations in Esophageal Squamous Cell Carcinoma. Am. J. Hum. Genet. 98, 256–274 (2016).
143. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. Nat. Genet. 45, 279–284 (2013).
144. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. Nature 524, 47–53 (2015).
145. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
146. Macintyre, G., Ylstra, B. & Brenton, J. D. Sequencing Structural Variants in Cancer for Precision Therapeutics. Trends Genet. 32, 530–542 (2016).
147. McClintock, B. The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics 26, 234–282 (1941).
148. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. 17, 1296–1303 (2007).
149. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467, 1109–1113 (2010).
150. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. Nature 578, 112–121 (2020).

151. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. Cell 153, 666–677 (2013).
152. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144, 27–40 (2011).
153. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. Cell 152, 1226–1236 (2013).
154. Kloosterman, W. P., Koster, J. & Molenaar, J. J. Prevalence and clinical implications of chromothripsis in cancer genomes. Curr. Opin. Oncol. 26, 64–72 (2014).
155. Kloosterman, W. P. *et al.* Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol. 12, R103 (2011).
156. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat. Genet. 52, 331–341 (2020).
157. Albertson, D. G. Gene amplification in cancer. Trends Genet. 22, 447–455 (2006).
158. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. Nat. Rev. Cancer 15, 371–381 (2015).
159. Merajver, S. D. *et al.* Germline BRCA1 mutations and loss of the wild-type allele in tumors from families with early onset breast and ovarian cancer. Clin. Cancer Res. 1, 539–544 (1995).
160. Affer, M. *et al.* Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. Leukemia 28, 1725–1735 (2014).
161. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. Nat. Genet. 50, 1262–1270 (2018).
162. Gajria, D. & Chandarlapaty, S. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. Expert Rev. Anticancer Ther. 11, 263–275 (2011).
163. Chan, K. C. A. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin. Chem. 59, 211–224 (2013).
164. Heitzer, E. *et al.* Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome Med. 5, 30 (2013).
165. McBride, D. J. *et al.* Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes Chromosomes Cancer 49, 1062–1069 (2010).
166. Leary, R. J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. Sci. Transl. Med. 2, 20ra14 (2010).
167. Mertens, F., Mandahl, N., Mitelman, F. & Heim, S. Cytogenetic analysis in the examination of solid tumors in children. Pediatr. Hematol. Oncol. 11, 361–377 (1994).
168. Markey, F. B., Ruezinsky, W., Tyagi, S. & Batish, M. Fusion FISH imaging: single-molecule detection of gene fusion transcripts in situ. PLoS One 9, e93488 (2014).
169. Ramkissoon, S. H. *et al.* Clinical implementation of integrated whole-genome copy number and mutation profiling for glioblastoma. Neuro. Oncol. 17, 1344–1355 (2015).

170. Davies, J. J., Wilson, I. M. & Lam, W. L. Array CGH technologies and their applications to cancer genomes. Chromosome Res. 13, 237–248 (2005).
171. Feldman, A. L. *et al.* Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. Blood 117, 915–919 (2011).
172. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 40, 722–729 (2008).
173. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. 77, 78–88 (2005).
174. Mardis, E. R. DNA sequencing technologies: 2006–2016. Nat. Protoc. 12, 213–218 (2017).
175. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. Nat. Rev. Genet. 21, 171–189 (2020).
176. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19, 329–346 (2018).
177. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351 (2016).
178. Ebbert, M. T. W. *et al.* Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20, 97 (2019).
179. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. 34, 303–311 (2016).
180. Spies, N. *et al.* Genome-wide reconstruction of complex structural variants using read clouds. Nat. Methods 14, 915–920 (2017).
181. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. Nat. Med. 26, 1271–1279 (2020).
182. Zhang, L. *et al.* Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell 181, 442–459.e29 (2020).
183. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345 (2018).
184. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162 (2019).
185. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138 (2009).
186. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. Genome Biol. 14, 405 (2013).
187. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524 (2016).
188. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 239 (2016).
189. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. Nat.

Methods 15, 201–206 (2018).
190. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305 (2019).
191. Aw, J. G. A. *et al.* Determination of isoform-specific RNA structure with nanopore long reads. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-0712-z.
192. Leger, A. *et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. bioRxiv (2020) doi:10.1101/843136.
193. Stephenson, W. *et al.* Direct detection of RNA modifications and structure using single molecule nanopore sequencing. bioRxiv (2020) doi:10.1101/2020.05.31.126763.
194. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. bioRxiv (2020) doi:10.1101/525741.
195. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017).
196. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608–611 (2015).
197. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. Nat. Methods 16, 88–94 (2019).
198. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. Nat. Biotechnol. 36, 321–323 (2018).
199. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. 10, 1784 (2019).
200. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 27, 677–685 (2017).
201. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. Cell 176, 663–675.e19 (2019).
202. Beyter, D. *et al.* Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. bioRxiv (2019) doi:10.1101/848366.
203. De Coster, W. *et al.* Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. 29, 1178–1187 (2019).
204. Cretu Stancu, M. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat. Commun. 8, 1326 (2017).
205. Tham, C. Y. *et al.* NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. 21, 56 (2020).
206. Mizuguchi, T. *et al.* A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J. Hum. Genet. 64, 359–368 (2019).
207. Merker, J. D. *et al.* Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine 20, 159–163 (2018).
208. Nattestad, M. *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. Genome Res. 28, 1126–1135

(2018).
209. Aganezov, S. *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. Genome Res. 30, 1258–1273 (2020).
210. Stangl, C. *et al.* Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. Nat. Commun. 11, 2861 (2020).
211. Gupta, A. *et al.* Single-molecule analysis reveals widespread structural variation in multiple myeloma. Proc. Natl. Acad. Sci. U. S. A. 112, 7689–7694 (2015).
212. Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat. Genet. 51, 1215–1221 (2019).
213. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-0538-8.
214. Arora, K. *et al.* Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. Sci. Rep. 9, 19123 (2019).
215. Fischerova, D., Zikan, M., Dundr, P. & Cibula, D. Diagnosis, Treatment, and Follow-Up of Borderline Ovarian Tumors. The Oncologist 17, 1515–1533 (2012).
216. Koshiyama, M., Matsumura, N. & Konishi, I. Recent concepts of ovarian carcinogenesis: type I and type II. Biomed Res. Int. 2014, 934261 (2014).
217. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. 45, 1127–1133 (2013).
218. Kurman, R. J. & Shih, I.-M. The Origin and Pathogenesis of Epithelial Ovarian Cancer: A Proposed Unifying Theory. The American Journal of Surgical Pathology 34, 433–443 (2010).
219. Thu, K. L. *et al.* A comprehensively characterized cell line panel highly representative of clinical ovarian high-grade serous carcinomas. Oncotarget 8, 50489–50499 (2017).
220. Fleury, H. *et al.* Novel high-grade serous epithelial ovarian cancer cell lines that reflect the molecular diversity of both the sporadic and hereditary disease. Genes & Cancer 6, 378–398 (2015).
221. Létourneau, I. J. *et al.* Derivation and characterization of matched cell lines from primary and recurrent serous ovarian cancer. BMC Cancer 12, 379 (2012).
222. Kreuzinger, C. *et al.* Molecular characterization of 7 new established cell lines from high grade serous ovarian cancer. Cancer Letters 362, 218–228 (2015).
223. Sachs, N. & Clevers, H. Organoid cultures for the analysis of cancer phenotypes. Current Opinion in Genetics & Development 24, 68–73 (2014).
224. Jones, P. M. & Drapkin, R. Modeling High-Grade Serous Carcinoma: How Converging Insights into Pathogenesis and Genetics are Driving Better Experimental Platforms. Front. Oncol. 3, 217 (2013).
225. Verissimo, C. S. *et al.* Targeting mutant RAS in patient-derived colorectal cancer organoids by combinatorial drug screening. Elife 5, (2016).
226. Hill, S. J. *et al.* Prediction of DNA Repair Inhibitor Response in Short-Term Patient-Derived Ovarian Cancer Organoids. Cancer Discovery 8, 1404–1421 (2018).

227. Gilmour, L. M. R. *et al.* Neuregulin expression, function, and signaling in human ovarian cancer cells. Clin. Cancer Res. 8, 3933–3942 (2002).
228. Aune, G. *et al.* Increased circulating hepatocyte growth factor (HGF): A marker of epithelial ovarian cancer and an indicator of poor prognosis. Gynecologic Oncology 121, 402–406 (2011).
229. Sheng, Q. *et al.* An Activated ErbB3/NRG1 Autocrine Loop Supports *In Vivo* Proliferation in Ovarian Cancer Cells. Cancer Cell 17, 298–310 (2010).
230. Bourgeois, D. L., Kabarowski, K. A., Porubsky, V. L. & Kreeger, P. K. High-grade serous ovarian cancer cell lines exhibit heterogeneous responses to growth factor stimulation. Cancer Cell Int. 15, 112 (2015).
231. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am. J. Hum. Genet. 72, 1117–1130 (2003).
232. Gabai-Kapara, E. *et al.* Population-based screening for breast and ovarian cancer risk due toBRCA1andBRCA2. Proceedings of the National Academy of Sciences 111, 14205–14210 (2014).
233. Wang, M. *et al.* PAX2 and PAX8 Reliably Distinguishes Ovarian Serous Tumors From Mucinous Tumors. Applied Immunohistochemistry & Molecular Morphology 23, 280–287 (2015).
234. Rajagopalan, H. & Lengauer, C. Aneuploidy and cancer. Nature 432, 338–341 (2004).
235. Gorringe, K. L. *et al.* High-Resolution Single Nucleotide Polymorphism Array Analysis of Epithelial Ovarian Cancer Reveals Numerous Microdeletions and Amplifications. Clinical Cancer Research 13, 4731–4739 (2007).
236. Hunter, S. M. *et al.* Pre-Invasive Ovarian Mucinous Tumors Are Characterized by CDKN2A and RAS Pathway Aberrations. Clinical Cancer Research 18, 5267–5277 (2012).
237. Romero, I., Sun, C. C., Wong, K. K., Bast, R. C. & Gershenson, D. M. Low-grade serous carcinoma: New concepts and emerging therapies. Gynecologic Oncology 130, 660–666 (2013).
238. Kuo, K.-T. *et al.* Analysis of DNA Copy Number Alterations in Ovarian Serous Tumors Identifies New Molecular Genetic Changes in Low-Grade and High-Grade Carcinomas. Cancer Research 69, 4036–4042 (2009).
239. Seidman, J. D., Yemelyanova, A., Zaino, R. J. & Kurman, R. J. The Fallopian Tube-Peritoneal Junction. International Journal of Gynecological Pathology 30, 4–11 (2011).
240. Kurman, R. J. & Shih, I.-M. Molecular pathogenesis and extraovarian origin of epithelial ovarian cancer—Shifting the paradigm. Human Pathology 42, 918–931 (2011).
241. Seidman, J. D. & Khedmati, F. Exploring the histogenesis of ovarian mucinous and transitional cell (Brenner) neoplasms and their relationship with Walthard cell nests: a study of 120 tumors. Arch. Pathol. Lab. Med. 132, 1753–1760 (2008).
242. Vassilev, L. T. *In Vivo* Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. Science 303, 844–848 (2004).

243. Yadav, B. *et al.* Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. Sci. Rep. 4, 5193 (2014).
244. Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. Science 355, 1152–1158 (2017).
245. Murai, J. Targeting DNA repair and replication stress in the treatment of ovarian cancer. International Journal of Clinical Oncology 22, 619–628 (2017).
246. Meijer, T. G. *et al.* Functional Ex Vivo Assay Reveals Homologous Recombination Deficiency in Breast Cancer Beyond BRCA Gene Defects. Clinical Cancer Research 24, 6277–6287 (2018).
247. Matano, M. *et al.* Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids. Nat. Med. 21, 256–262 (2015).
248. Fumagalli, A. *et al.* Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. Proceedings of the National Academy of Sciences 114, E2357–E2364 (2017).
249. Schmeler, K. M. *et al.* Neoadjuvant chemotherapy for low-grade serous carcinoma of the ovary or peritoneum. Gynecol. Oncol. 108, 510–514 (2008).
250. Gershenson, D. M. *et al.* Recurrent low-grade serous ovarian carcinoma is relatively chemoresistant. Gynecologic Oncology 114, 48–52 (2009).
251. Pectasides, D. *et al.* Advanced stage mucinous epithelial ovarian cancer: The Hellenic Cooperative Oncology Group experience. Gynecologic Oncology 97, 436–441 (2005).
252. Brown, J. & Frumovitz, M. Mucinous tumors of the ovary: current thoughts on diagnosis and management. Curr. Oncol. Rep. 16, 389 (2014).
253. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (2013).
254. Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43, (2013).
255. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297–1303 (2010).
256. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
257. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22, 568–576 (2012).
258. Garrison, E. & Mart, G. E. Haplotype-based variant detection from short-read sequencing. arXiv (2012).
259. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
260. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6, 80–92 (2012).
261. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using

next-generation sequencing data. Bioinformatics 28, 423–425 (2012).
262. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222 (2016).
263. Consortium, T. G. of T. N. & The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825 (2014).
264. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015).
265. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Systems 3, 385–394.e3 (2016).
266. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
267. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
268. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
269. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191 (2009).
270. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
271. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369 (2014).
272. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8, 2281–2308 (2013).
273. Koo, B.-K. *et al.* Controlled gene expression in primary Lgr5 organoid cultures. Nat. Methods 9, 81–83 (2012).
274. Timmermans, M., Sonke, G. S., Van de Vijver, K. K., van der Aa, M. A. & Kruitwagen, R. F. P. M. No improvement in long-term survival for epithelial ovarian cancer patients: A population-based study between 1989 and 2014 in the Netherlands. Eur. J. Cancer 88, 31–37 (2018).
275. Mirza, M. R. *et al.* Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. N. Engl. J. Med. 375, 2154–2164 (2016).
276. Bleijs, M., van de Wetering, M., Clevers, H. & Drost, J. Xenograft and organoid model systems in cancer research. EMBO J. 38, e101654 (2019).
277. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 141, 1762–1772 (2011).
278. Kopper, O. *et al.* An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity. Nat. Med. 25, 838–849 (2019).
279. Hill, S. J. *et al.* Prediction of DNA Repair Inhibitor Response in Short-Term Patient-Derived Ovarian Cancer Organoids. Cancer Discov. 8, 1404–1421 (2018).

280. Jabs, J. *et al.* Screening drug effects in patient-derived cancer cells links organoid responses to genome alterations. Mol. Syst. Biol. 13, 955 (2017).
281. Ooft, S. N. *et al.* Patient-derived organoids can predict response to chemotherapy in metastatic colorectal cancer patients. Sci. Transl. Med. 11, (2019).
282. Swan, H. A. *et al.* Abstract 1619: Personalized medicine: A CLIA-certified high-throughput drug screening platform for ovarian cancer. Cancer Res 13 Supplement, (2018).
283. Hoogstraat, M. *et al.* Genomic and transcriptomic plasticity in treatment-naive ovarian cancer. Genome Res. 24, 200–211 (2014).
284. Cameron, D. L. *et al.* GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv (2019) doi:10.1101/781013.
285. Böhm, S. *et al.* Chemotherapy Response Score: Development and Validation of a System to Quantify Histopathologic Response to Neoadjuvant Chemotherapy in Tubo-Ovarian High-Grade Serous Carcinoma. J. Clin. Oncol. 33, 2457–2463 (2015).
286. Rustin, G. J. S. *et al.* Re: New guidelines to evaluate the response to treatment in solid tumors (ovarian cancer). Journal of the National Cancer Institute vol. 96 487–488 (2004).
287. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer 45, 228–247 (2009).
288. Schumacher, D. *et al.* Heterogeneous pathway activation and drug response modelled in colorectal-tumor-derived 3D cultures. PLoS Genet. 15, e1008076 (2019).
289. Yan, H. H. N. *et al.* A Comprehensive Human Gastric Cancer Organoid Biobank Captures Tumor Subtype Heterogeneity and Enables Therapeutic Screening. Cell Stem Cell 23, 882–897.e11 (2018).
290. Nguyen, L., Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. bioRxiv (2020) doi:10.1101/2020.01.13.905026.
291. Cotto, K. C. *et al.* DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. 46, D1068–D1073 (2018).
292. Heilmann, A. M. *et al.* CDK4/6 and IGF1 receptor inhibitors synergize to suppress the growth of p16INK4A-deficient pancreatic cancers. Cancer Res. 74, 3947–3958 (2014).
293. Rosato, R. R. *et al.* Mechanism and functional role of XIAP and Mcl-1 down-regulation in flavopiridol/vorinostat antileukemic interactions. Mol. Cancer Ther. 6, 692–702 (2007).
294. Smith, G. *et al.* Activating K-Ras mutations outwith 'hotspot' codons in sporadic colorectal tumours - implications for personalised cancer medicine. Br. J. Cancer 102, 693–703 (2010).
295. Cooke, A. Biochemical and Biological Characterization of KRAS Q61 Mutants. (University of North Carolina at Chapel Hill, 2018).
296. Samimi, G., Katano, K., Holzer, A. K., Safaei, R. & Howell, S. B. Modulation of the cellular pharmacology of cisplatin and its analogs by the copper exporters ATP7A and ATP7B. Mol. Pharmacol. 66, 25–32 (2004).
297. Samimi, G. *et al.* Increased expression of the copper efflux transporter ATP7A mediates resistance to cisplatin, carboplatin, and oxaliplatin in ovarian cancer cells. Clin. Cancer Res.

10, 4661–4669 (2004).
298. Nakayama, K. *et al.* Copper-transporting P-type adenosine triphosphatase (ATP7B) as a cisplatin based chemoresistance marker in ovarian carcinoma: comparative analysis with expression of MDR1, MRP1, MRP2, LRP and BCRP. Int. J. Cancer 101, 488–495 (2002).
299. Hegde, G. V. *et al.* Blocking NRG1 and other ligand-mediated Her4 signaling enhances the magnitude and duration of the chemotherapeutic response of non-small cell lung cancer. Sci. Transl. Med. 5, 171ra18 (2013).
300. Shang, Z.-F. *et al.* 4E-BP1 participates in maintaining spindle integrity and genomic stability via interacting with PLK1. Cell Cycle 11, 3463–3471 (2012).
301. Del Bufalo, D. *et al.* Endothelin-1 protects ovarian carcinoma cells against paclitaxel-induced apoptosis: requirement for Akt activation. Mol. Pharmacol. 61, 524–532 (2002).
302. Liu, R. *et al.* The Akt-specific inhibitor MK2206 selectively inhibits thyroid cancer cells harboring mutations that can activate the PI3K/Akt pathway. J. Clin. Endocrinol. Metab. 96, E577–85 (2011).
303. Aqeilan, R. I. *et al.* Targeted deletion of Wwox reveals a tumor suppressor function. Proc. Natl. Acad. Sci. U. S. A. 104, 3949–3954 (2007).
304. Schirmer, M. A. *et al.* Relevance of Sp Binding Site Polymorphism in WWOX for Treatment Outcome in Pancreatic Cancer. J. Natl. Cancer Inst. 108, (2016).
305. Bunn, P. A., Jr *et al.* Expression of Her-2/neu in human lung cancer cell lines by immunohistochemistry and fluorescence in situ hybridization and its relationship to *in vitro* cytotoxicity by trastuzumab and chemotherapeutic agents. Clin. Cancer Res. 7, 3239–3250 (2001).
306. Kimura, K. *et al.* Antitumor effect of trastuzumab for pancreatic cancer with high HER-2 expression and enhancement of effect by combined therapy with gemcitabine. Clin. Cancer Res. 12, 4925–4932 (2006).
307. Stankova, J., Shang, J. & Rozen, R. Antisense inhibition of methylenetetrahydrofolate reductase reduces cancer cell survival *in vitro* and tumor growth *in vivo*. Clin. Cancer Res. 11, 2047–2052 (2005).
308. McCluggage, W. G. *et al.* Data set for reporting of ovary, fallopian tube and primary peritoneal carcinoma: recommendations from the International Collaboration on Cancer Reporting (ICCR). Mod. Pathol. 28, 1101–1122 (2015).
309. Phan, N. *et al.* A simple high-throughput approach identifies actionable drug sensitivities in patient-derived tumor organoids. Commun Biol 2, 78 (2019).
310. Maru, Y., Tanaka, N., Itami, M. & Hippo, Y. Efficient use of patient-derived organoids as a preclinical model for gynecologic tumors. Gynecol. Oncol. 154, 189–198 (2019).
311. Gotimer, K., Chen, H., Leiserowitz, G. S. & Smith, L. H. Short-term organoid culture for drug sensitivity testing in high-grade serous ovarian cancer. Gynecol. Oncol. 154, 92–93 (2019).
312. Tiriac, H. *et al.* Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. Cancer Discov. 8, 1112–1129 (2018).
313. Saleem, A. & Price, P. M. Early tumor drug pharmacokinetics is influenced by tumor perfu-

sion but not plasma drug exposure. Clin. Cancer Res. 14, 8184–8190 (2008).
314. Pujol, J. L. *et al.* Tumor-tissue and plasma concentrations of platinum during chemotherapy of non-small-cell lung cancer patients. Cancer Chemother. Pharmacol. 27, 72–75 (1990).
315. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
316. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv (2018) doi:10.1101/201178.
317. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
318. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 27, 2050–2060 (2017).
319. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28, 423–425 (2012).
320. Yang, L. *et al.* Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes. Cell 157, 1736 (2014).
321. Zhang, C.-Z. *et al.* Chromothripsis from DNA damage in micronuclei. Nature 522, 179–184 (2015).
322. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. Cell 163, 1641–1654 (2015).
323. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. Nat. Rev. Genet. 19, 453–467 (2018).
324. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer 7, 233–245 (2007).
325. Mansfield, A. S. *et al.* Neoantigenic Potential of Complex Chromosomal Rearrangements in Mesothelioma. J. Thorac. Oncol. 14, 276–287 (2019).
326. de Vree, P. J. P. *et al.* Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat. Biotechnol. 32, 1019–1025 (2014).
327. Hillmer, A. M. *et al.* Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res. 21, 665–675 (2011).
328. Sanders, A. D. *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. Nat. Biotechnol. 38, 343–354 (2020).
329. Greer, S. U. *et al.* Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. Genome Med. 9, 57 (2017).
330. Gong, T., Hayes, V. M. & Chan, E. K. F. Detection of somatic structural variants from short-read next-generation sequencing data. Brief. Bioinform. (2020) doi:10.1093/bib/bbaa056.
331. Lee, A. Y. *et al.* Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. Genome Biol. 19, 188 (2018).
332. Sarwal, V. *et al.* A comprehensive benchmarking of WGS-based structural variant callers.

bioRxiv (2020) doi:10.1101/2020.04.16.045120.
333. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196 (2010).
334. Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. Sci. Rep. 6, 24607 (2016).
335. Velazquez-Villarreal, E. I. *et al.* Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. Commun Biol 3, 318 (2020).
336. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. 6, 10001 (2015).
337. Cameron, D. L. *et al.* GRIDSS2: harnessing the power of phasing and single breakends in somatic structural variant detection. bioRxiv (2020) doi:10.1101/2020.07.09.196527.
338. Fujimoto, A. *et al.* Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res. (2020) doi:10.1101/gr.255026.119.
339. Andersson, A. K. *et al.* The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. Nat. Genet. 47, 330–337 (2015).
340. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (2013).
341. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
342. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468 (2018).
343. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 8, 14061 (2017).
344. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
345. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034 (2015).
346. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
347. Xi, R., Lee, S., Xia, Y., Kim, T.-M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. 44, 6274–6286 (2016).
348. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. Nat. Methods 12, 1058–1060 (2015).
349. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115 (2012).
350. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222 (2016).

351. Belyeu, J. R. *et al.* SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. Gigascience 7, (2018).
352. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34 (2017).
353. Turkbey, B., Pinto, P. A. & Choyke, P. L. Imaging techniques for prostate cancer: implications for focal therapy. Nat. Rev. Urol. 6, 191–203 (2009).
354. Gerwing, M. *et al.* The beginning of the end for conventional RECIST — novel therapies require novel imaging approaches. Nat. Rev. Clin. Oncol. 16, 442–458 (2019).
355. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat. Rev. Genet. 20, 71–88 (2019).
356. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. Nat. Rev. Cancer 11, 426–437 (2011).
357. Klega, K. *et al.* Detection of Somatic Structural Variants Enables Quantification and Characterization of Circulating Tumor DNA in Children With Solid Tumors. JCO Precis Oncol 2018, (2018).
358. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. 50, 1388–1398 (2018).
359. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat. Commun. 10, 3240 (2019).
360. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20, 117 (2019).
361. Husain, H. *et al.* Cell-Free DNA from Ascites and Pleural Effusions: Molecular Insights into Genomic Aberrations and Disease Biology. Mol. Cancer Ther. 16, 948–955 (2017).
362. Harris, F. R. *et al.* Quantification of Somatic Chromosomal Rearrangements in Circulating Cell-Free DNA from Ovarian Cancers. Sci. Rep. 6, 29831 (2016).
363. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat. Biotechnol. 38, 433–438 (2020).
364. Quick, J. Ultra-long read sequencing protocol for RAD004 v3 (protocols.io.mrxc57n). doi:10.17504/protocols.io.mrxc57n.
365. van Dessel, L. F. *et al.* The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. Nat. Commun. 10, 5251 (2019).
366. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573–580 (1999).
367. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 47, D853–D858 (2019).
368. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. Science 297, 1003–1007 (2002).

369. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods 12, 780–786 (2015).
370. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. Nature 526, 68–74 (2015).
371. Korenchuk, S. *et al.* VCaP, a cell-based model system of human prostate cancer. *In Vivo* 15, 163–168 (2001).
372. Middelkamp, S. *et al.* Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants. Genome Med. 11, 79 (2019).
373. van Dessel, L. F. *et al.* Application of circulating tumor DNA in prospective clinical oncology trials - standardization of preanalytical conditions. Mol. Oncol. 11, 295–304 (2017).
374. van Dessel, L. F. *et al.* High-throughput isolation of circulating tumor DNA: a comparison of automated platforms. Mol. Oncol. 13, 392–402 (2019).
375. Kaplan, J. A. Leukemia in Children. Pediatr. Rev. 40, 319–331 (2019).
376. Dupain, C., Harttrampf, A. C., Urbinati, G., Geoerger, B. & Massaad-Massade, L. Relevance of Fusion Genes in Pediatric Cancers: Toward Precision Medicine. Mol. Ther. Nucleic Acids 6, 315–326 (2017).
377. Kuhlen, M., Klusmann, J.-H. & Hoell, J. I. Molecular Approaches to Treating Pediatric Leukemias. Front Pediatr 7, 368 (2019).
378. Campana, D. & Behm, F. G. Immunophenotyping of leukemia. J. Immunol. Methods 243, 59–75 (2000).
379. Jongen-Lavrencic, M. *et al.* Molecular Minimal Residual Disease in Acute Myeloid Leukemia. N. Engl. J. Med. 378, 1189–1199 (2018).
380. Paietta, E. Assessing minimal residual disease (MRD) in leukemia: a changing definition and concept? Bone Marrow Transplant. 29, 459–465 (2002).
381. Campana, D. Minimal residual disease in acute lymphoblastic leukemia. Semin. Hematol. 46, 100–106 (2009).
382. Modvig, S. *et al.* Minimal residual disease quantification by flow cytometry provides reliable risk stratification in T-cell acute lymphoblastic leukemia. Leukemia 33, 1324–1336 (2019).
383. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood 127, 2391–2405 (2016).
384. De Braekeleer, E. *et al.* ABL1 fusion genes in hematological malignancies: a review. Eur. J. Haematol. 86, 361–371 (2011).
385. D'Angiò, M. *et al.* Clinical features and outcome of SIL/TAL1-positive T-cell acute lymphoblastic leukemia in children and adolescents: a 10-year experience of the AIEOP group. Haematologica 100, e10–3 (2015).
386. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. 23, 227–238.e3 (2018).
387. Meyer, C. *et al.* Human MLL/KMT2A gene exhibits a second breakpoint cluster region for recurrent MLL-USP2 fusions. Leukemia 33, 2306–2340 (2019).

388. Winters, A. C. & Bernt, K. M. MLL-Rearranged Leukemias-An Update on Science and Clinical Approaches. Front Pediatr 5, 4 (2017).
389. Pongers-Willemse, M. J. *et al.* Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets Report of the BIOMED-1 CONCERTED ACTION: Investigation of minimal residual disease in acute leukemia. Leukemia 13, 110–118 (1999).
390. Cooper, M. D. & Alder, M. N. The Evolution of Adaptive Immune Systems. Cell vol. 124 815–822 (2006).
391. Immunobiology: the immune system in health and disease. Choice Reviews Online vol. 32 32–3876 (1995).
392. Schatz, D. G. & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. Nature Reviews Immunology vol. 11 251–263 (2011).
393. Roth, D. B. V(D)J Recombination: Mechanism, Errors, and Fidelity. Mobile DNA III 311–324 (2015).
394. Jeske, D. J., Jarvis, J., Milstein, C. & Capra, J. D. Junctional diversity is essential to antibody activity. J. Immunol. 133, 1090–1092 (1984).
395. Szczepański, T. *et al.* Comparative analysis of Ig and TCR gene rearrangements at diagnosis and at relapse of childhood precursor-B–ALL provides improved strategies for selection of stable PCR targets for monitoring of minimal residual disease. Blood 99, 2315–2323 (2002).
396. van Dongen, J. J. M., van der Velden, V. H. J., Brüggemann, M. & Orfao, A. Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood 125, 3996–4009 (2015).
397. van Dongen, J. J. *et al.* Standardized RT-PCR analysis of fusion gene transcripts from chromosome aberrations in acute leukemia for detection of minimal residual disease. Report of the BIOMED-1 Concerted Action: investigation of minimal residual disease in acute leukemia. Leukemia 13, 1901–1928 (1999).
398. van der Velden, V. H. J. *et al.* Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. Leukemia 17, 1013–1034 (2003).
399. Sánchez, R., Ayala, R. & Martínez-López, J. Minimal Residual Disease Monitoring with Next-Generation Sequencing Methodologies in Hematological Malignancies. Int. J. Mol. Sci. 20, (2019).
400. Leukemia and Lymphoma Society (LLS): Minimal residual disease. https://www.lls.org/sites/default/files/National/USA/Pdf/Publications/FS35_MRD_Final_2019.pdf.
401. Mason, J. & Griffiths, M. Detection of Minimal Residual Disease in Leukaemia by RT-PCR. Methods in Molecular Biology 269–280 (2011) doi:10.1007/978-1-60761-947-5_18.
402. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98 (2011).

403. Stangl, C. *et al.* Partner-independent fusion gene detection by multiplexed CRISPR/Cas9 enrichment and long-read Nanopore sequencing. Nat Commun (2020) doi:10.1101/807545.
404. Janssen, J. W., Ludwig, W. D., Sterry, W. & Bartram, C. R. SIL-TAL1 deletion in T-cell acute lymphoblastic leukemia. Leukemia 7, 1204–1210 (1993).
405. Robinson, J. T. *et al.* Integrative genomics viewer. Nat. Biotechnol 29, 24–26 (2011).
406. IMGT Repertoire (IG and TR) 1. Locus and genes. http://www.imgt.org/IMGTrepertoire/LocusGenes/#C.
407. Langerak, A. W. *et al.* Unraveling the consecutive recombination events in the human IGK locus. J. Immunol. 173, 3878–3888 (2004).
408. Custom Alt-R® CRISPR-Cas9 guide RNA. Integrative DNA Technologies. https://www.idtdna.com/site/order/designtool/index/CRISPR_CUSTOM.
409. Valle-Inclan, J. E. *et al.* Rapid identification of genomic structural variations with nanopore sequencing enables blood-based cancer monitoring. medRxiv (2019) doi:10.1101/19011932.
410. van der Velden, V. H. J. & van Dongen, J. J. M. MRD detection in acute lymphoblastic leukemia patients using Ig/TCR gene rearrangements as targets for real-time quantitative PCR. Methods Mol. Biol. 538, 115–150 (2009).
411. IMGT Repertoire (IG and TR) 1. Locus and genes. http://www.imgt.org/IMGTrepertoire/LocusGenes/#C.
412. Singh, M. *et al.* High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. 10, 3120 (2019).
413. Mehravar, M., Shirazi, A., Mehrazar, M. M. & Nazari, M. Pre-validation of Gene Editing by CRISPR/Cas9 Ribonucleoprotein. Avicenna J. Med. Biotechnol. 11, 259–263 (2019).
414. Rusk, N. More accurate nanopore sequencing. Nature methods vol. 16 460 (2019).
415. New research algorithms yield accuracy gains for nanopore sequencing. Oxford Nanopore Technologies http://nanoporetech.com/about-us/news/new-research-algorithms-yield-accuracy-gains-nanopore-sequencing (2020).
416. IMGT/BlastSearch. http://www.imgt.org/BlastSearch/.
417. Wang, Y., Wu, N., Liu, D. & Jin, Y. Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. Curr. Genomics 18, 378 (2017).
418. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
419. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746 (2017).
420. Chen, H. *et al.* Short-term organoid culture for drug sensitivity testing of high-grade serous carcinoma. Gynecol. Oncol. 157, 783–792 (2020).
421. Nguyen, L., W M Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. Nat. Commun. 11, 5584 (2020).
422. Blank, C. U., Haanen, J. B., Ribas, A. & Schumacher, T. N. The 'cancer immunogram'. Science 352, 658–660 (2016).

423. Dijkstra, K. K. *et al.* Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. Cell 174, 1586–1598.e12 (2018).
424. Cattaneo, C. M. *et al.* Tumor organoid–T-cell coculture systems. Nat. Protoc. 15, 15–39 (2020).
425. Thibodeau, M. L. *et al.* Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. Genet. Med. 22, 1892–1897 (2020).
426. Eichler, E. *et al.* The structure, function, and evolution of a complete human chromosome 8. bioRxiv (2020) doi:10.21203/rs.3.rs-72559/v1.
427. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84 (2020).
428. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. Nature 581, 444–451 (2020).
429. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-00746-x.
430. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat. Biotechnol. (2020) doi:10.1038/s41587-020-0731-9.
431. Miller, D. E. *et al.* Targeted long-read sequencing resolves complex structural variants and identifies missing disease-causing variants. bioRxiv (2020) doi:10.1101/2020.11.03.365395.

SUMMARY

Cancer genomics is a thriving field with constant methodological and technological advances. These developments enable personalized and targeted treatments for cancer patients based on the unique genomic profiles of tumors. In this thesis I leveraged novel technologies for the advancement of cancer research and care. In **chapter 1** I introduced the different types of genomic variation, the genomic characteristics of cancer, the importance of genomic technology for personalized medicine and the potential of liquid biopsies for low-invasive cancer monitoring. I also introduced model systems used in cancer research, including patient-derived organoids (PDOs), detailing the potential for their use particularly in ovarian cancer (OC). Lastly, I introduced the role of somatic SV in cancer and the different sequencing technologies that are used to detect them, along with the challenges that this presents.

In the first part of this thesis, we used organoid technology to advance OC research. In **chapter 2** we established and characterized a PDO biobank that faithfully represented the disease, and presented its applications. The biobank consisted of 56 PDO lines from 32 patients and included all the main histological subtypes of OC. We characterized the organoid lines at the histopathological, genomic and epigenomic levels and showed how these lines faithfully recapitulated the original tumors. We also demonstrated that these characteristics were maintained over time and extensive culturing. Furthermore, transcriptomic analysis showed that the PDO lines from the different OC subtypes clustered together based on expression patterns. Intra- and inter-patient heterogeneity was represented in the biobank. We also presented proof-of-concept for drug screening studies with these PDOs, with different subtype responses to platinum-based chemotherapy. Resistance to chemotherapy in recurrent disease was also reflected in the biobank. Lastly, we showed that the PDOs could be xenografted into mice for *in vivo* drug-sensitivity assays and that PDOs were amenable to genomic modification to study tumorigenesis (for tumorigenesis study). Overall, our results highlighted the potential of organoids to advance OC research.

Next, we expanded the OC-PDO biobank and used those PDOs for extensive drug screening in **chapter 3**. We performed screening assays on 36 PDO lines derived from 23 patients and retrospectively compared their drug responses to clinical responses of the patients. Evaluating several clinical biomarkers, we showed that PDOs recapitulated patient response to chemotherapy and targeted treatments. I.e. the most responsive PDOs in drug screening assays had been derived from patients that had better clinical response to the same treatment. Furthermore, inter- and intra-patient drug response heterogeneity was found between different PDO lines and could be partially explained by genomic

aberrations. Lastly, we were able to obtain drug response results within three weeks of tumor sampling, illustrating the short turnaround time needed for these assays, a critical feature for clinical implementation. Overall, these results highlighted the importance of deriving PDO lines from multiple tumor locations of a patient, when applicable, to improve drug response prediction and clinical decision making based on PDO assays.

In the second part of the thesis I focused on somatic structural variation in cancer. Accurate detection of structural variants (SVs) is still challenging, and truth sets and standardized workflows are lacking. We tackled the challenge of accurate somatic SV detection in cancer genomes and generated a truth set of somatic SVs that can be used for method development and benchmarking, which I presented in **chapter 4**. We performed genome-wide analysis of the paired melanoma and normal lymphoblastoid COLO829 cell lines. We sequenced these cell lines to a high depth using four different sequencing technologies: Illumina, Oxford Nanopore, Pacific Biosciences and 10X Genomics. We also performed experimental validation, including Bionano optical mapping data, to ensure that the truth set was reliable and complete. Finally, we demonstrated the utility of the truth set by determining the SV detection performance of each technology as a function of tumor purity and sequencing depth. This gold-standard truth set, together with the underlying multi-platform genomic characterization of this cancer cell line pair, are an important resource for benchmarking and method development efforts in the cancer genomics community.

Furthermore, we developed methods to utilize long-read sequencing and somatic structural variants (SVs) for cancer dynamics after treatment and minimal residual disease tracing. In **chapter 5**, we developed an assay that leveraged nanopore sequencing technology for rapid detection of somatic SVs from a tumor. We used low-coverage whole-genome sequencing of a tumor with nanopore technology and then used filtering and random-forest classification to select the most likely somatic SV candidates. We could retrieve these biomarker-candidates within three days after tissue obtention. These somatic SVs could be used, after validation, as patient-specific biomarkers for cancer tracking in circulating tumor DNA (ctDNA) by ultra-sensitive PCR methods. We applied our assay to ten ovarian and prostate cancer samples and obtained multiple biomarkers per sample in mere days. We demonstrated retrospectively that longitudinal monitoring of cancer dynamics was feasible using these somatic SV biomarkers. Summarizing, our method enabled rapid and cost-effective identification of a set of patient-specific SVs that can be used to study ctDNA dynamics.

We also developed an assay that leverages CRISPR-Cas9 based enrichment of genomic targets in pediatric leukemias from the lymphoid lineage. In **chapter 6**, we targeted

loci recurrently involved in genomic rearrangements in these leukemias, such as the immunoglobulin (Ig) and T-cell receptor (TCR) loci, and the KMT2A and SIL-TAL1 fusion-gene loci. These loci are widely used as PCR-based minimal residual disease (MRD) tracing based on the patient-specific rearrangements. We applied our assay to ten acute lymphoid leukemia samples and showed that we successfully enriched the fusion, Ig and TCR loci. We successfully validated known MRD targets in these patients within two days after sample obtention and identified an additional set of patient-specific rearrangements. Our approach poses as an attractive alternative to current multi-step biomarker identification assays in lymphoid leukemia, with increased speed and detection sensitivity.

Finally, in **chapter 7** I discussed and reflected on the technological advances presented in the previous chapters. I explained the advantages of PDO technology in OC research, but also the challenges for its further clinical implementation in clinical care. Similarly, I identify current challenges and propose several solutions for enhancing the knowledge of the role of somatic SVs in cancer and implementation of long-read sequencing. In conclusion, this thesis proposes several cancer genomics technological opportunities to advance cancer research and develop personalized diagnostic assays to improve patients' outcomes.

SAMENVATTING

Onderzoek naar het kanker genoom is dynamisch en gaat gepaard met constante vooruitgang op zowel methodologisch als technologisch vlak. Deze ontwikkelingen maken gepersonaliseerde en gerichte behandelingen voor kankerpatiënten mogelijk op basis van de unieke genomische profielen van tumoren. Voor dit proefschrift heb ik gebruik gemaakt van nieuwe technologieën voor de verbetering van kankeronderzoek en -zorg. In **hoofdstuk 1** heb ik de verschillende soorten genomische variatie, de genomische kenmerken van kanker, het belang van genomische technologie voor gepersonaliseerde geneeskunde en het potentieel van vloeibare biopsieën voor minimaal invasieve kankermonitoring geïntroduceerd. Daarnaast heb ik modelsystemen geïntroduceerd die worden gebruikt bij kankeronderzoek, waaronder PDO's (patiënt-afgeleide organoïden), waarin het potentieel voor hun gebruik, met name bij eierstokkanker wordt beschreven. Ten slotte heb ik de rol van somatische structurele variaties (SVs) bij kanker geïntroduceerd en de verschillende sequencing technieken die worden gebruikt om ze te detecteren, samen met de uitdagingen die dit met zich meebrengt.

In het eerste deel van dit proefschrift hebben we organoïden technologie gebruikt om onderzoek naar eierstokkanker te bevorderen. In **hoofdstuk 2** hebben we een PDO biobank opgezet en de toepassingen ervan gepresenteerd. De biobank bestond uit 56 PDO-lijnen van 32 patiënten en omvatte alle belangrijke histologische subtypes van eierstokkanker. We hebben de organoïde lijnen gekarakteriseerd op histopathologisch, genomisch en epigenomisch niveau en lieten zien hoe deze lijnen nauwgezet de oorspronkelijke tumoren nabootsten. We hebben ook aangetoond dat deze kenmerken na langdurige kweek behouden bleven. Verder toonde analyse van het transcriptoom aan dat de PDO-lijnen van de verschillende subtypen van eierstokkanker samen clusteren op basis van expressiepatronen. Ook was de heterogeniteit binnen en tussen patiënten vertegenwoordigd in de biobank. We presenteerden tevens een proof-of-concept voor drug screeningsonderzoeken op deze PDO's met verschillende subtype-reacties op platina-bevattende chemotherapie. Resistentie tegen chemotherapie bij recidief kwam ook tot uiting in de biobank. Ten slotte hebben we aangetoond dat de PDO's konden worden getransplanteerd in muizen voor in vivo medicatie gevoeligheids assays en dat PDO's bruikbaar waren voor genetische modificatie voor onderzoek naar tumorgenese. Onze resultaten tonen het potentieel van organoïden om eierstokkankeronderzoek te bevorderen.

In **hoofdstuk 3** hebben we de PDO biobank van eierstokkanker uitgebreid en de PDO's gebruikt voor uitgebreide drugsscreening. We hebben screenings uitgevoerd op 36 PDO-lijnen die zijn afgeleid van 23 patiënten en we hebben retrospectief de reactie van

de PDO-lijnen op medicatie vergeleken met de klinische respons van de patiënten. Door verschillende klinische biomarkers te evalueren, hebben we aangetoond dat PDO's de respons van de patiënt op chemotherapie en gerichte behandelingen nabootsen. De meest responsieve PDO's in de screenings voor medicatie waren afkomstig van patiënten die een betere klinische respons hadden op dezelfde behandeling. Bovendien werd inter- en intra-patiënt heterogeniteit van medicatie respons aangetoond tussen verschillende PDO-lijnen en kon deze gedeeltelijk worden verklaard door genomische afwijkingen. Tot slot waren we in staat om binnen drie weken na het afnemen van tumorweefsel resultaten te verkrijgen over de respons op medicatie, wat de korte doorlooptijd die nodig is voor deze assays illustreert, een cruciaal kenmerk voor klinische implementatie. Al met al benadrukken deze resultaten het belang van het verkrijgen van PDO-lijnen van meerdere tumorlocaties van een patiënt om de voorspelling van medicatie respons en klinische besluitvorming op basis van PDO assays te verbeteren.
In het tweede deel van het proefschrift heb ik me gericht op somatische structurele variatie in kanker. Nauwkeurige detectie van SV's is nog steeds een uitdaging, en vastgestelde referentiesets en gestandaardiseerde workflows ontbreken. We zijn de uitdaging van nauwkeurige somatische SV-detectie in kankergenomen aangegaan en hebben een referentieset van somatische SV's gegenereerd die kan worden gebruikt voor het ontwikkelen van methoden en benchmarking, wat ik in **hoofdstuk 4** heb besproken. We hebben een genoom-brede analyse uitgevoerd van de gepaarde melanoom- en normale lymfoblastoïde COLO829-cellijnen. We hebben deze cellijnen met diepe coverage gesequenced met behulp van vier verschillende sequencing technieken: Illumina, Oxford Nanopore, Pacific Biosciences en 10X Genomics. Ook hebben we experimentele validatie uitgevoerd, inclusief Bionano optische mapping, om ervoor te zorgen dat de referentieset betrouwbaar en volledig was. Ten slotte hebben we het nut van de referentieset aangetoond door de SV-detectie prestaties van elke technologie te bepalen afhankelijk van de zuiverheid van de tumor en de sequencing diepte. Deze referentieset, samen met de onderliggende multi-platform genomische karakterisering van dit kankercellijnenpaar, is een belangrijke bron voor benchmarking en methodologische ontwikkeling in het genomics veld.

Daarnaast hebben we methoden ontwikkeld om 'long-read' sequencing en somatische SV's te gebruiken om de dynamiek van kanker na behandeling te bestuderen en om minimale restziekten op te sporen. In **hoofdstuk 5** hebben we een assay ontwikkeld die gebruikmaakt van Nanopore sequencing technologie voor snelle detectie van somatische SV's in een tumor. We hebben met lage coverage een tumor gesequenced met Nanopore technologie en hebben filtering en random-forest-classificatie toegepast om de meest waarschijnlijke somatische SV-kandidaten te selecteren. We konden deze biomarker-kandidaten binnen drie dagen na ontvangst van het weefsel terugvinden. Deze

somatische SV's zouden, na validatie, kunnen worden gebruikt als patiënt-specifieke biomarkers voor het volgen van kanker in circulerend tumor-DNA (ctDNA) door middel van ultragevoelige PCR-methoden. We hebben onze assay toegepast op tien ovarium- en prostaatkankermonsters en hebben in slechts enkele dagen meerdere biomarkers per monster verkregen. We hebben in retrospect aangetoond dat longitudinale monitoring van de kankerdynamiek mogelijk was met behulp van deze somatische SV-biomarkers. Samenvattend maakte onze methode een snelle en kosteneffectieve identificatie mogelijk van een reeks patiëntspecifieke SV's die kunnen worden gebruikt om ctDNA-dynamiek te bestuderen.

We hebben ook een assay ontwikkeld die gebruik maakt van op CRISPR-Cas9 gebaseerde verrijking van genomische targets bij pediatrische leukemieën uit de lymfoïde lijn. In **hoofdstuk 6** hebben we ons gericht op loci die herhaaldelijk betrokken zijn bij genomische herschikkingen bij deze leukemieën, zoals de immunoglobuline (Ig) en T-celreceptor (TCR) loci, en de KMT2A en SIL-TAL1 fusiegenen loci. Deze loci worden veel gebruikt voor op PCR gebaseerde tracering van minimale restziekte op basis van de patiëntspecifieke herschikkingen. We hebben onze assay toegepast op tien acute lymfoïde leukemie monsters en hebben aangetoond dat we met succes de fusie, Ig en TCR loci hebben verrijkt. We hebben met succes bekende doelen van minimale restziekte bij deze patiënten gevalideerd binnen twee dagen na ontvangst van het monster en we hebben een aanvullende reeks patiëntspecifieke herschikkingen geïdentificeerd. Onze aanpak vormt een geschikt alternatief voor de huidige meerstaps biomarker-identificatie assays bij lymfoïde leukemie, met verhoogde snelheid en detectiegevoeligheid.

Ten slotte heb ik in **hoofdstuk 7** de technologische vooruitgang die in de vorige hoofdstukken zijn gepresenteerd bediscussieerd. Ik heb de voordelen van PDO-technologie voor eierstokkankeronderzoek uitgelegd, maar ook de uitdagingen voor de verdere implementatie in de klinische zorg. Daarnaast besprak ik de bestaande uitdagingen omtrent somatische SV's en stel ik verschillende oplossingen voor om de kennis van de rol van somatische SV's bij kanker en implementatie van Nanopore sequencing technologie te vergroten. Concluderend stelt dit proefschrift verschillende kanker genomics gebaseerde technologische mogelijkheden voor om kankeronderzoek vooruit te helpen en gepersonaliseerde diagnostische assays te ontwikkelen om het ziektebeloop van patiënten positief te beïnvloeden.

ACKNOWLEDGEMENTS

What an accomplishment to write a thesis! But even though only my name appears on the cover, many more people are responsible for this success. Some of their names can be read throughout the chapters but others have had a more indirect but equally relevant role. I have tried to compile all those names here. However, there are so many that the fear of forgetting a name (never a person) is real. So, if you feel that you should be here, then you definitely are and shame on me. Thank you so much!

I want to start by thanking Wigard for giving me the opportunity to start my PhD and for all the enthusiasm and ideas that started all the projects that I have written here. Edwin, thank you for taking me into your group for the last part of the PhD and for all the help in the last push. Thank you to the members of my doctoral examination committee, prof. Susanne Lens, prof. Emile Voest, dr. Jeroen de Ridder, prof. Hans Kristian Ploos van Amstel and prof. Tobias Marschall for taking the time to critically read this thesis. Thank you to prof. Ronald Zweemer and prof. Els Witteveen for their guidance through the years and for their different and very necessary clinical point of view. Also thank you to dr. Jan Molenaar for his guidance through the PhD supervision committee.

To the triples, Chris and Christina. I simply would not have made it without you. Apart from your obvious contribution to most chapters in this thesis, thank you for always being there to carry me through the difficult times (mostly winter and when back from holiday). Thank you for all the coffees, walks, tears, cakes, breakfasts, lunches, dinners, games, skiing, sailing, travel surprises and everything else that we have had and will continue having together. Chris, thank you for your kindness and positive energy. I have learned a lot from you about life, gynaecology and plenty of interesting medical facts. Looking forward to the day you party with us. Stay yellow! Christina, thank you for upgrading me from colleague to friend so quickly (even though there was a short regression at some point). Thank you for carrying me through with much needed deadlines (seriously), thank you for many lasagnas and some vodka-snows, but mainly for your contagious kindness and serenity that I keep learning from you. Extended thank yous to Rob, Victor (and Pau!) for forming the quintuples, hopefully we can all meet again soon, Meanwhile, take care of these two!

To the Wizards, it was really a shame that our group fell apart. Thank you for all the rotating activities and dinners and Thursday morning meetings (with mandatory cookies!). We even went up the Alpe d'Huez together! Mircea, thank you for showing me around the UMC and around Utrecht. We shared many coffees, (long) lunch breaks, random conversations, retreats, beers and shots and I missed all of it after you left. Mark, thank

you for helping and teaching me so much and always with a smile and a joke. I really thought I would not manage when you left the UMC. Ivo, thank you for all the knowledge about beer and whisky, and for generating almost all the data I have used. Glen, thank you for all the support and glentoring, and for literally carrying me up a mountain (twice!). Ellen, thank you for your enthusiasm, support and mentoring and for always being there for a late evening venting chat. Joline, thank you for keeping up with all the OC meetings and for your clinical point of view that brought a lot into the group. Alessio, thank you for your passion and for an unforgettable amount of beer at Regensburg. Joe, thank you for your eye for detail and special point of view, I hope someday we can celebrate that house warming. Sam, thanks for teaching me how to be a supervisor and for all your help.

To the Cuppen group, Ewart, Arne, Sharon, Bastiaan, Nicolle, Lissane, Sjors, Robin, Ies, Sander, Luan, thank you for adopting me for the last year of my PhD. It was not the smoothest time and you all helped to make it better. Keep up the great work but more importantly keep having fun every day. Special thanks to my officemates: Judith, thank you for all the late PhD venting and for translating the samenvatting. I am sure you will have a very successful end of your PhD. Roel, thank you for a lot of random news and threads that started with speciation, many puns (most of them were good), for your help and for being my open-source guru.

Thank you to all the colleagues from the Genetics department and the CMM for all the retreats, masterclasses, (theme) borrels, sorely missed through 2020, and overall fun and healthy work environment. There are too many people to mention them all but if you feel included in this acknowledgement it is because you are. Miguel, thank you for all the (long) lunch and coffee breaks, keeping me updated with the gossip and intense ping pong games on Fridays. Joske, thank you for your enthusiasm, you were definitely a major contributor to the fun mentioned above. You are brilliant and any group is lucky to have you, even remotely. Also thank you to all the workers that make things run smoothly in our daily work life, starting with Monique and Cristina, always there to solve all the daily PhD student headaches, but also including HPC and IT, cleaning and horeca staff.

Thank you to all the collaborators that made this thesis possible. To Oded and Kadi from the Hubrecht Institute for all their organoid work, it was a pleasure working with you. To Anouk and Martijn from the Erasmus MC, thank you for believing in our nanopore idea and making it a reality with enthusiasm.

Thank you to all the friends and teammates from the Utrecht Bulls that have shared more than a basketball court over these years. Maikel, Arne, Tommy, Maurice, Joost, Michiel,

Davey, Jitse, Jordy, Giancarlo, Maurice, Nick, Peter, JJ and many more, thank you for all the practices, games, drinks, team weekends (even abroad), activities, chicken, MITs, bacamangos, and fortunately or unfortunately a whole lot of online gaming. We may not have won much but we for sure had quite some fun. Thank you Redmer, for all of the above and on top of it being an awesome housemate (and unexpected officemate) during the hard last stages of this thesis.

A los fundadores de Catania, gracias por conseguir que Holanda empezase a sentirse un hogar. Pau y Fre, por intensos juegos de Catan y esos 5 de mayo. Miguel y Lorenza, gracias por ser únicos y un apoyo siempre y por impulsar aquel viaje inolvidable a Costa Rica. Emilio, Lord, gracias por estar siempre dispuesto a cualquier aventura, por haberte caído de un árbol en Málaga y por todas las distracciones con diversos juegos. Andrés, Mae, gracias por todas las noches que pasamos en La Haya, por la aventura Malagasy y por enseñarnos a hablar en tico. Prepárate porque cualquier día nos presentamos todos en Costa Rica para aprender a surfear.

Gracias a todos los Coraçaos, "el grupo", que ya sabéis quienes sois. Porque no sería quien soy sin vosotros y aunque la distancia pasa factura siempre estáis conmigo. En particular a dos compañeros de andanzas académicas: Sergio, gracias por las llamadas de (des)motivación, la aventura en Madagascar y la visita que algún día haré a Berlín (aunque ya no estés). Y Nacho, gracias por estar siempre disponible, y por su sentido común. También gracias a Carmen, porque pese a todo aprendí mucho contigo y de ti. Muito obrigado à Joana por todas as aventuras deste ano de covid e muitas mais que virão. Tu é(s) e tu sabes.

Gracias a todos mis tíos, tías, primos y primas. A mis abuelos, estén donde estén estarán muy orgullosos, y sobre todo a mis abuelas, Ángeles y Pepa que sufren la distancia más que nadie, gracias por siempre creer a ciegas en mí y por vuestro ejemplo de bondad sin límite. Os quiero muchísimo. A mi hermana Claudia, gracias por enseñarme a tener paciencia y por conocerme mejor que nadie. Y, finalmente y más que a nadie, gracias a mi Pai y a mi Mai, por apoyarnos y darnos todo lo que hemos necesitado. Siempre me dejásteis hacer y no es fácil, aunque de momento sigue bien. Os debo todo.

LIST OF PUBLICATIONS

PART OF THIS THESIS

Kopper O, de Witte CJ*, Lõhmussaar K*, Valle-Inclan JE*, Hami N, Kester L, *et al.* **An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity.** Nature Medicine 25, 838–849 (2019).
https://doi.org/10.1038/s41591-019-0422-6

de Witte CJ, Valle-Inclan JE, Hami N, Lõhmussaar K, Kopper O, Vreuls CPH, *et al.* **Patient-Derived Ovarian Cancer Organoids Mimic Clinical Response and Exhibit Heterogeneous Inter- and Intrapatient Drug Responses.** Cell Reports 31, 11 (2020).
https://doi.org/10.1016/j.celrep.2020.107762

Valle-Inclan JE, Besselink N, Renkens I, van Roosmalen MJ, Hastie A, Kwint M, van Lieshout S, Nelen M, Pang A, Priestley P, Wenger A, Ylstra B, Fijneman RJA, Kloosterman WP, Cuppen E. **A multi-platform reference for somatic structural variation detection.** Submitted and available as a preprint at bioRxiv 340497 (2020).
https://doi.org/10.1101/2020.10.15.340497

Valle-Inclan JE*, Stangl C*, de Jong AC, van Dessel LF, van Roosmalen MJ, Helmijr JCA, *et al.* **Rapid identification of genomic structural variations with nanopore sequencing enables blood-based cancer monitoring.** Submitted and available as a preprint at medRxiv 19011932 (2019).
https://doi.org/10.1101/19011932

OTHER PUBLICATIONS

Stangl C, de Blank S, Renkens I, Westera L, Verbeek T, Valle-Inclan JE, *et al.* **Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing.** Nature Communications 11, 2861 (2020).
https://doi.org/10.1038/s41467-020-16641-7

Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Valle-Inclan JE, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP. **Mapping and phasing of structural variation in patient genomes using nanopore sequencing.** Nature Com-

munications 8, 1326 (2017).
https://doi.org/10.1038/s41467-017-01343-4

Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, Cuppen E, *et al.* **GRIDSS2: harnessing the power of phasing and single breakends in somatic structural variant detection.** Submitted and available at bioRxiv 196527 (2020).
https://doi.org/10.1101/2020.07.09.196527

Datema E, Hulzink RJM, Blommers L, Valle-Inclan JE, van Orsouw N, Wittenberg AHJ, *et al.* **The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only**. Preprint available at bioRxiv 084772 (2016).
https://doi.org/10.1101/084772

Cook DE*, Valle-Inclan JE*, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. **Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing.** Plant Physiology 179 (1) 38-54 (2019)
https://doi.org/10.1104/pp.18.00848

CURRICULUM VITAE

Jose Antonio Espejo Valle-Inclan was born on March 4th 1992 in Sevilla, Spain. He grew up in Madrid with his parents and younger sister. He attended primary and high school at Instituto Ramiro de Maeztu in Madrid, where he also discovered a passion for basketball. After graduating with honors in 2010, he started his Bachelor of Science in Biochemistry at the Universidad Autonoma de Madrid. In the last year of those studies, in 2014, he followed bioinformatics and programming elective courses, which prompted him to move countries and start a Master of Science in Bioinformatics at Wageningen University, the Netherlands. During his master thesis and internship, he started working with genome and transcriptome assembly and nanopore sequencing technology in the field of plant pathology. In 2016 he moved to Utrecht to start his PhD at the University Medical Center Utrecht under the supervision of dr. Wigard Kloosterman and prof. Edwin Cuppen, of which this thesis is the final product. He will continue his career as a postdoctoral fellow at the European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK.