nature genetics

Article

https://doi.org/10.1038/s41588-022-01230-9

The genomic and immune landscape of long-term survivors of high-grade serous ovarian cancer

In the format provided by the authors and unedited

Supplementary Note

CONTENTS

1. Supplementary Results

1.1 Patients

Inclusion criteria were FIGO Stage IIIC/IV, high-grade serous cancer (HGSC, confirmed by histopathology review), ovarian, fallopian tube or primary peritoneal carcinoma, meeting survival group criteria; either short-term survivors (death within 2 years); moderate-term survivors (death ≥2 and <10 years); or long-term survivors (overall survival ≥ 10 years; Extended Data Fig. 1a). Age at diagnosis, tumor grade and FIGO stage were similar between survival groups (Extended Data Fig. 1b, Supplementary Table 1). All cases were classified HGSC according to the current two tier WHO classification[1] for serous ovarian cancer, and there was no difference between the survival groups according to the Silverberg three tier ovarian cancer grading system[2] that was previously in common use. Long-term survivors were more likely to have no macroscopic residual disease following primary debulking surgery (13/60, 22%) compared with short-term (1/34, 3%) and moderate-term survivors (2/32, 6%; P = 0.0227; Extended Data Fig. 1b).

Compared to short-term survivors, moderate and long-term survivors were less likely to have primary peritoneal cancer (P = 0.1010) and neoadjuvant chemotherapy (P = 0.0640), although neither feature was significantly different between the groups (Extended Data Fig. 1b). We note that patients who received neoadjuvant chemotherapy were all either short- (n = 3) or moderate-term survivors (n = 1), and therefore may have unique molecular or clinical features. However, the four neoadjuvant cases did not belong to unique mutational signature clusters (FDR adjusted P value (P_{adj}) = 0.6131, Extended Data Fig. 4b), methylation clusters (P_{adj} = 0.0654, Supplementary Fig. 11) or immune clusters (P_{adj} = 0.8976, Extended Data Fig. 9a), and there were no significant associations between molecular or clinical features in neoadjuvant compared to adjuvant cases, aside from progression-free and overall survival (data not shown).

1.2 Homologous recombination deficiency analysis

The functional impact of homologous recombination gene alterations was measured in tumor genomes using the Classifier of Homologous Recombination Deficiency (CHORD)[3] method, which integrates base substitution, indel, and structural rearrangement signatures to generate *BRCA1*-type homologous recombination deficiency (HRD) and *BRCA2*-type HRD prediction scores. The majority of *BRCA1* germline mutant (19/24, 79.2%), *BRCA1* promoter methylated (14/16, 87.5%) and *BRCA1* somatic mutant (8/10, 80%) tumor genomes were classified as

BRCA1-type HRD (*BRCA1*-type CHORD score >0.5; Fig. 1b). All *BRCA2* germline mutant (11/11, 100%) and *BRCA2* somatic mutant (5/5, 100%) tumors were classified as *BRCA2*-type HRD (*BRCA2*-type CHORD score >0.5; Fig. 1b).

We noted that while some *BRCA1* altered tumors had a *BRCA1*-type CHORD score less than 0.5 (9/50, 18%), most of these (7/9, 77.8%) had a score between 0.15 and 0.5 that was consistently higher than wild-type homologous recombination proficient tumors (mean CHORD score of 0.01, range 0.00 to 0.04), indicating that the 0.5 cut-off may exclude some *BRCA1* altered tumors. Of the two *BRCA1* altered tumors with an exceptionally low *BRCA1*-type CHORD score, AOCS-079 (*BRCA1*-type CHORD score 0.02) had a somatic *BRCA1* frameshift mutation in exon 10, and MMAY00631 (*BRCA1*-type CHORD score 0.03) had a germline missense mutation (NM_007294.3:c.5096G>A), which is considered pathogenic but with a reduced penetrance compared to the average truncating *BRCA1* mutation[4].

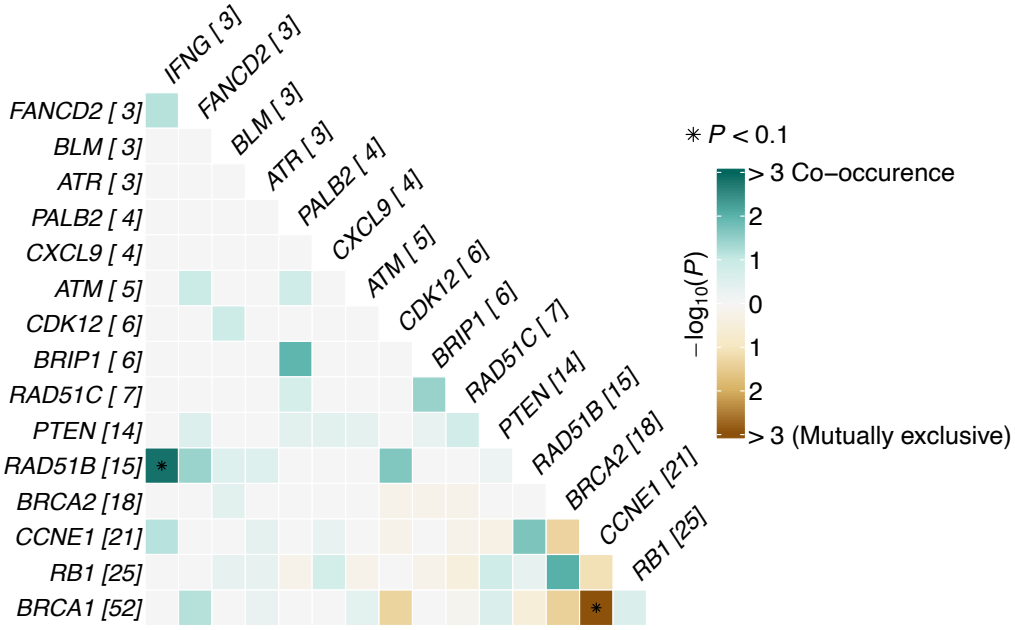
All tumors in patients with germline mutations in *RAD51C* (n = 2), *PALB2* (n = 1) and *BRIP1* (n = 2) were classified as *BRCA2*-type HRD (*BRCA2*-type CHORD score >0.5; Fig. 1b), indicating a shared mutational process with *BRCA2* mutated tumors. Of the tumors with somatic homologous recombination gene alterations (other than *BRCA1* and *BRCA2*), 62.5% (5/8) were classified as *BRCA2*-type HRD, including tumors driven by *RAD51D* deletion (n = 1), *RAD51C* promoter methylation (n = 2), *RAD51C* splice site mutation (n = 1), and *PTEN* deletion (n = 1). The three tumors with somatic homologous recombination gene alterations that were classified as homologous recombination proficient (CHORD score of <0.01), included one *ATM* structural variant (SV), one *RAD51B* SV, and one *RAD51C* SV.

CDK12 inactivation has been previously associated with HRD[5] and PARP inhibitor sensitivity[6], however, we found that tumors with somatic *CDK12* mutations demonstrated low *BRCA1*-type and *BRCA2*-type HRD signatures (CHORD score <0.01; Fig. 1b), and were thus classified as homologous recombination proficient. Rather, *CDK12* inactivated tumors were characterized by tandem duplications (Fig. 1b), as previously shown in HGSC[7] and metastatic prostate cancer[8]. Notably, although there were only six cases with *CDK12* mutations, 83.3% (5/6) were long-term survivors.

Retention of the wild-type germline allele of *BRCA1* or *BRCA2* in HGSC tumor genomes has been associated with treatment resistance and short survival[9], however we found that all germline homologous recombination pathway mutations (40/40, 100%) were associated with loss of the wild-type allele (Supplementary Table 5).

CCNE1 amplification tends to be mutually exclusive with *BRCA1* and *BRCA2* mutations[10,11], and thus is a marker for homologous recombination proficiency. Indeed 90.1%

(19/21) of *CCNE1* amplified tumors were classified as homologous recombination proficient (CHORD score <0.5; Fig. 1b). Of the two *CCNE1* amplified tumors classified as HRD, one had a germline *BRIP1* mutation (*BRCA2*-type HRD) and the other had a somatic *BRCA1* SV (*BRCA1*-type HRD). Accordingly, *CCNE1* amplification tended to be mutually exclusive with *BRCA1* alterations (P_{adj} = 0.0169, Fisher's exact; co-occurrence in one case, 0.79%; Supplementary Fig. 1, Supplementary Table 15). *CCNE1* amplification and *BRCA2* alterations did not co-occur in any primary tumor samples, however this gene pair did not reach statistical significance (P_{adj} = 0.4554; Supplementary Fig. 1, Supplementary Table 15).



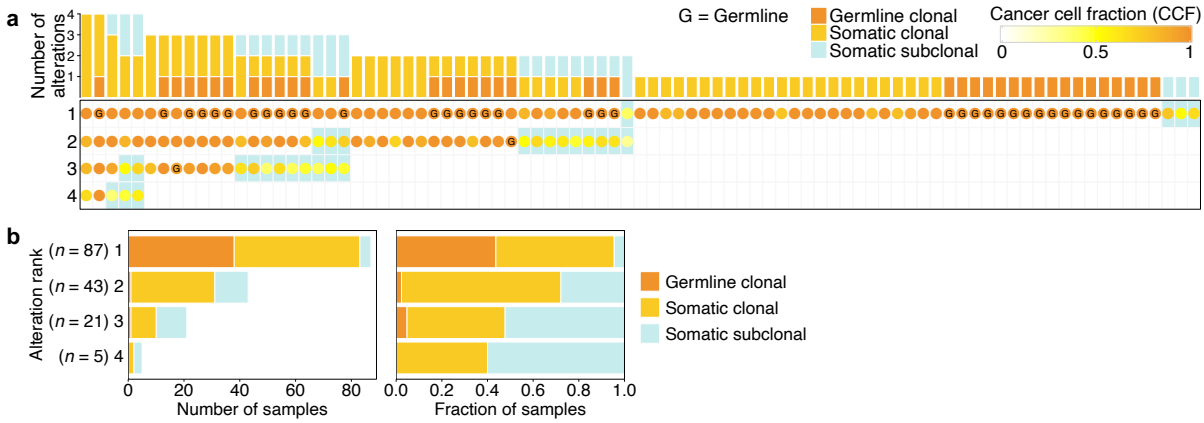
Supplementary Fig. 1 | Mutual exclusivity and co-occurrence of DNA repair pathway alterations. The degree of overlap (co-occurrence) or mutual exclusivity of germline and somatic alterations in genes of interest was measured in primary tumors (n = 126). Genes were those listed in Fig. 1b (i.e. homologous recombination pathway genes, other DNA repair genes, *CCNE1*, and immune genes) with alterations in at least three primary tumors. Each square represents the relationship between two genes, the first gene is labelled on the left (rows) and the second gene is labelled at the top (columns). To avoid testing mutually dependent observations, *CXCL9* amplification was considered representative of the 4q13 amplicon, which also includes the genes *CXCL10* and *CXCL11*. The number of samples with an alteration is indicated in parentheses after each gene name. The color of each square indicates the mutual exclusivity or co-occurrence of a gene pair, with the intensity of the color indicating the -$\log_{10}$

P value. *P* values (two-sided) were calculated using the Fisher's exact test, with multiple hypothesis testing correction applied (* $P_{adj} < 0.1$).

In the remaining subset of tumors with no *CCNE1* amplification, *CDK12* mutations or homologous recombination gene alterations identified (i.e. homologous recombination wild-type), 91.3% (21/23) were classified as homologous recombination proficient (CHORD score <0.5; Fig. 1b). Two of these cases (MMAY00566 and MAOC02193) were classified by CHORD as HRD with no underlying homologous recombination gene alteration identified, both of which were long-term survivors.

Some tumor genomes had evidence of a mixture of both *BRCA1*-type and *BRCA2*- type signatures (Fig. 1b). Using a CHORD probability score cut-off of >0.15 for both *BRCA1*-type and *BRCA2*-type HRD, 14 tumors had an apparent contribution of both HRD signatures, which in some cases could be attributed to independent alterations in at least two homologous recombination genes. For example, *BRCA1* germline mutation case AOCS-143 also had somatic *RAD51C* promoter methylation, and evidence of both *BRCA1*- and *BRCA2*-type CHORD scores (*BRCA1*-type 0.64, *BRCA2*-type 0.20). In another example, *BRCA2* germline mutation carrier MAOC00673 also had a somatic inter-chromosomal translocation disrupting *BRCA1*, and evidence of both HRD signatures (*BRCA2*-type 0.64, *BRCA1*-type 0.28).

To predict whether DNA repair pathway alterations tended to be clonal (present in all tumor cells) or subclonal (present in a fraction of tumor cells), we performed subclonal reconstruction estimation using bulk tumor DNA sequence data (see Supplementary Methods). While we were unable to assess clonality of *BRCA1* and *RAD51C* promoter methylation events, all remaining germline and somatic SNVs, indels, structural variants and copy number variants in homologous recombination and DNA repair genes were included in this analysis. As expected, all germline alterations (100%, 40/40) were classified as clonal, and there was a mixture of both clonal and subclonal somatic alterations (Supplementary Fig. 2a).



Supplementary Fig. 2 | Clonality of DNA repair pathway alterations. a, Bars at the top represent the number of DNA repair pathway alterations (y axis) within a sample (excluding promoter methylation of *BRCA1* and *RAD51C*), and bar colors indicate the alteration classification as either germline clonal, somatic clonal or somatic subclonal. The heatmap below shows the cancer cell fraction (CCF) of each alteration, with germline alterations labelled "G" and cells containing a subclonal mutation are highlighted blue. Rows are ranked 1 to 4 based on highest to lowest CCF within a sample. Columns (tumor samples) are sorted first by the total number of alterations, then by the total somatic clonal mutations and finally by the number of germline clonal mutations. Primary tumor DNA sample IDs are labelled across the bottom of each column. **b**, Bars show the number (left) and proportion (right) of samples with DNA repair pathway alterations classified as germline clonal, somatic clonal or somatic subclonal, grouped by the number of DNA repair alterations detected (samples with at least 1 alteration n = 87, at least 2 alterations n = 43, at least 3 alterations n = 21, and 4 alterations n = 5).

For samples with multiple DNA repair pathway alterations, the alterations were ranked 1 to 4 from highest to lowest cancer cell fraction (CCF), which is an estimation of the fraction of cancer cells in which a given mutation is present[12]. In samples with at least two alterations (n = 43), the majority of second hits were clonal (72.1%, 31/43; Supplementary Fig. 2b). In tumor samples with more than two DNA repair alterations, the third and fourth alterations were more likely to be subclonal, with 52.4% (11/21) and 60% (3/5) classified as subclonal respectively (Supplementary Fig. 2b).

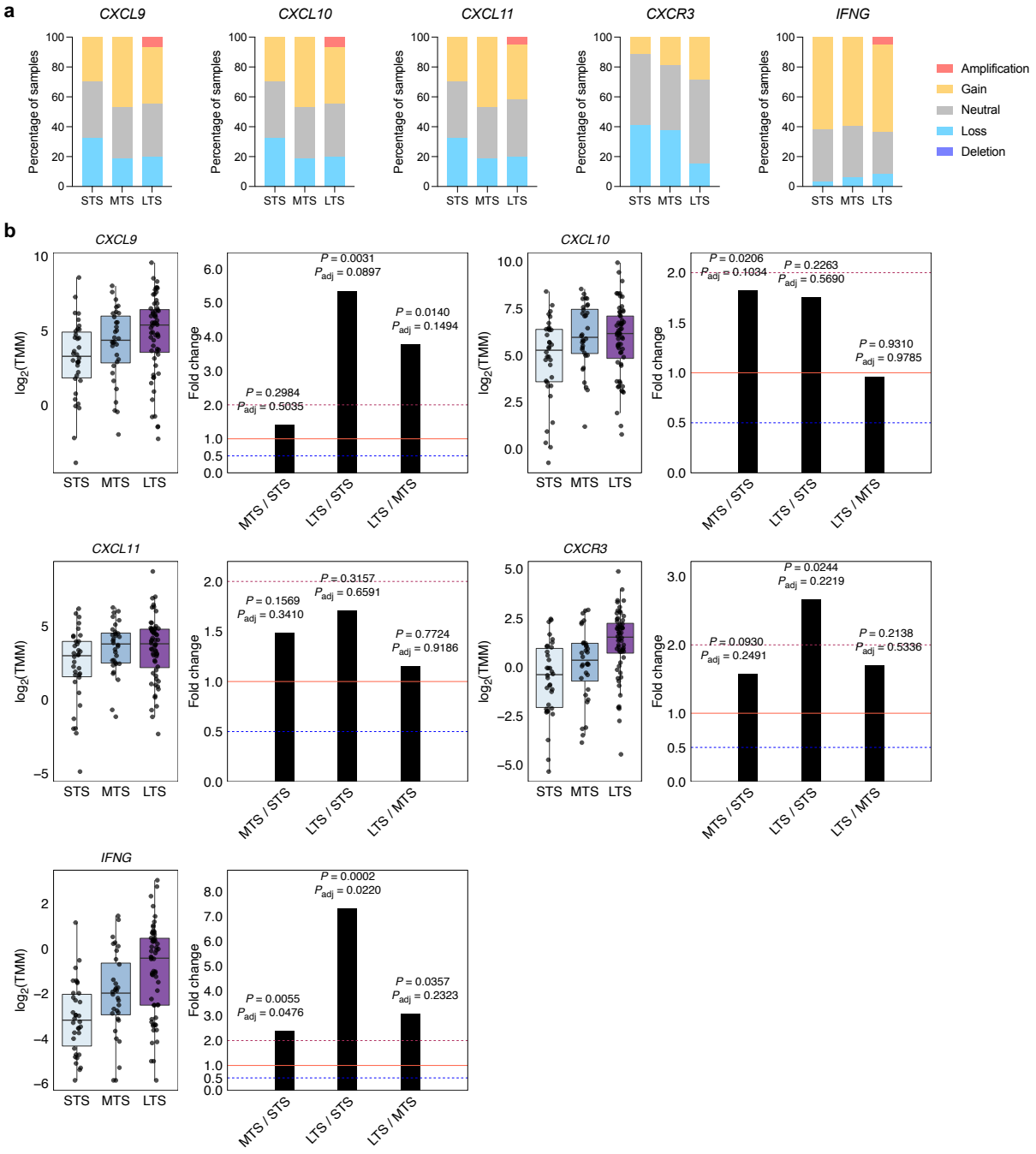
1.3 Recurrent gene mutation analysis

As is characteristic of HGSC[13], non-synonymous *TP53* mutations were present in 100% (126/126) of primary tumor samples (Fig. 3, Supplementary Table 7). Confirming previous reports[10,14], apart from *TP53*, somatic point mutations and indels in cancer-associated genes are rare, with only two other genes statistically enriched with non-synonymous SNVs and indels: *BRCA1* and *RB1* ($P_{adj} < 0.1$; Extended Data Fig. 2a, Supplementary Data 1).

Consistent with previous observations in HGSC[14,15], structural rearrangements that disrupt tumor suppressor genes were more common than point mutations (Extended Data Fig. 2a). In addition to known putative driver SVs in *RB1*, *NF1*, *PTEN* and *RAD51B*[14], this analysis identified recurrent disruption of the tumor suppressor genes *MYH9*, *EZH2*, *ARID1B*, *TBL1XR1*, *ARID1A*, *YWHAE*, *CREBBP*, *RHOA*, *ATRX*, *AXIN1* and *STAG1* ($P_{adj} < 0.1$; Extended Data Fig. 2a; Supplementary Data 2).

Copy number analysis revealed extensive somatic copy number alterations (SCNA) across the cohort of 126 primary tumors, including 20 amplified and 47 deleted genomic regions ($P_{adj} < 0.1$; Supplementary Data 3). Common, frequent SCNA regions included amplification of 8q24 (61.1%, 77 of 126 samples), 3q26 (52.4%, 66/126) and 12p13 (23%, 29/126), containing known ovarian cancer-associated genes *MYC*, *MECOM* and *CCND2* respectively, and loss of 6q27 (74.6%, 94/126), 13q14 (70.6%, 89/126), and 10q23 (34.1%, 43/126), containing *ARID1B*, *RB1* and *PTEN* respectively.

Amplification of 4q13 and 12q15 was only observed in tumors of long-term survivors (Fig. 1b). Notably, both regions contain genes that encode cytokines (4q13: *CXCL9*, *CXCL10*, and *CXCL11*, and 12q15: *IFNG*). Although not subject to homozygous deletion, copy loss of *CXCR3* at Xq.13, which encodes the receptor for CXCL9, CXCL10 and CXCL11, was frequent in short-term survivors (41.2%, 14/34) and moderate-term survivors (37.5%, 12/32) compared to long-term survivors (15%, 9/60; Supplementary Fig. 3a). Analysis of transcriptome data showed higher expression of cytokines in long-term survivors relative to short-term survivors, with a significant increase seen in *IFNG* expression ($P_{adj} = 0.0220$) (Supplementary Fig. 3b). In an independent HGSC cohort (The Cancer Genome Atlas, TCGA[10,16]) high mRNA expression of *CXCL11* and *IFNG* was associated with better survival ($P < 0.05$; Supplementary Table 8), with a trend towards better survival observed in patients with high expression of *CXCL9* ($P = 0.052$) and *CXCL10* ($P = 0.081$), which have previously been associated with improved survival in HGSC patients[17,18].



Supplementary Fig. 3 | Copy number and transcriptional alterations of cytokines. a, Proportion of copy number events in genes encoding cytokines (*IFNG*, *CXCL9*, *CXCL10*, *CXCL11*) and the receptor *CXCR3* across survival groups (n = 34 short-term survivors, STS; n = 32 moderate-term survivors, MTS; n = 60 long-term survivors, LTS). **b**, Boxplots (left) summarize gene expression by survival group (STS n = 34, MTS n = 32, LTS n = 60); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, and whiskers show the smallest/largest values within 1.5 times the interquartile range. Differential expression analysis was performed using DESeq2 to determine

fold change (right) of gene expression between survival groups (two-tailed Wald test, unadjusted *P* values and Benjamini-Hochberg adjusted *P* values (P_{adj}) are shown).

To investigate whether observed genomic alterations were associated with survival, all high confidence non-synonymous SNVs, indels, SVs and copy number alterations (restricted to homozygous deletions and amplifications) affecting known cancer-associated genes were tallied, and the frequencies compared across survival groups (Extended Data Fig. 2). Alterations were enriched in several cancer-associated genes (*EXT1*, *YWHAE*, *RAD21*, *ZNRF3*, *BLM*, and *ALK*) in long-term survivors relative to short-term survivors, while alterations in other cancer-associated genes (*CEP89*, *CEBPA*, *CCNE1*, and *KTN1*) were enriched in short-term survivors relative to long-term survivors (Extended Data Fig. 2; $P < 0.05$, Fisher's test). For example, *CEP89* had a mutation frequency of 32.4% (11/34) in short-term survivors, mostly due to gene amplification, compared to 6.3% (2/32) in moderate-term survivors and 8.3% (5/60) of long-term survivors (Extended Data Fig. 2). Accordingly, high expression of *CEP89* in TCGA was associated with poor survival ($P = 0.0084$; Supplementary Table 8). The mutational frequency of an additional 8 cancer-associated genes were significantly different between either moderate- and short-term survivors (*RECQL4*, *UBR5*, *MYC*, *CSMD3*, and *NDRG1*), or long- and moderate-term survivors (*ANK1*, *PPARG*, and *SCG2*; Extended Data Fig. 2; $P < 0.05$, Fisher's test).

RHOA is a cancer related gene that was flagged as frequently affected by SVs (GRIN analysis), as well as in the GISTIC deletion peak at 3p21 ($P_{adj} < 0.0001$). Combined mutation analysis indicated the mutation frequency was not significantly different between survival groups, however high expression of this gene correlates with longer survival in TCGA[10] ($P = 0.032$). Further investigation is warranted to determine if *RHOA* deletion is the key driver of the deletion peak at 3p21, or if indeed other nearby gene targets may be important. Long-term survivors had the highest frequency of *RHOA* gain (46.7%, 28/60) compared to short-term (38.2%, 13/34) and moderate-term survivors (28.1%, 9/32), which may partially explain the survival association with expression observed in TCGA. It should be noted that *RHOA* is considered both an oncogene and a tumor suppressor gene in the Cancer Gene Census[19].

1.4 Gene fusions

A total of 42,705 gene fusion events were detected in RNA-seq data of 126 primary tumors and 3 relapse tumors using Arriba[20] (Supplementary Data 5), and of these 555 were in-frame

fusion events with medium or high confidence (according to Arriba annotations). Seven gene fusions were recurrent (detected in more than one patient), and 2 were considered genuine gene fusions upon manual review (excluding tandem duplications, read-through events, and variants associated with poor mapping quality). These were *USP7-CARHSP1* (16p13.2) and *KIF1B-PGD* (1p36.22), each with a frequency of 2 out of 126 primary tumors (1.6%). These findings are consistent with our previous report that recurrent in-frame gene fusions are rare in HGSC[14]. Neither *USP7-CARHSP1* or *KIF1B-PGD* fusions have been observed in previous HGSC studies[21-23] or in other cancers[24]. The two *USP7-CARHSP1* fusions were both detected in long-term survivors and the *KIF1B-PGD* fusions were found in one long-term and one moderate-term survivor.

1.5 Relapse samples

To examine relapse disease in long-term survivors we were able to whole-genome sequence five samples from four patients collected during additional surgeries (Fig. 2a). We confirmed that all samples collected at relapse were clonally related to primary tumors, for example sharing *TP53* mutations (Supplementary Table 7, Fig. 2c) and somatic SVs (Fig. 2b). Three of the relapse patients carried *BRCA1* mutations, for which there was no evidence of secondary reversion mutations at recurrence, and the *BRCA1*-type CHORD scores were consistent between primary and relapse tumor samples (Fig. 2c). All relapse tumors had a higher mutation burden and SV burden than primary tumors (Fig. 2c), likely associated with further genomic evolution and exposure to DNA damaging therapeutic agents.

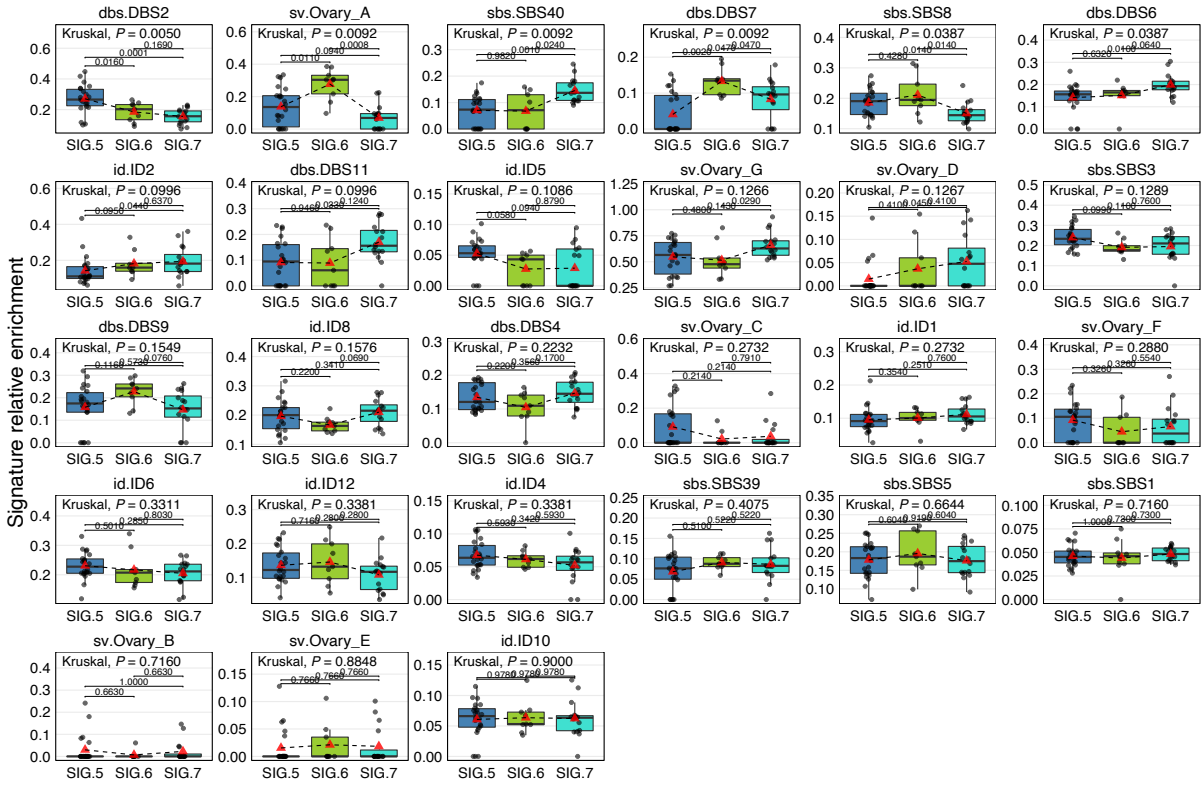
Patient MWMH00552 had the lowest proportion of shared SVs between primary and relapse tumor samples and the longest period of disease-free remission (13.5 years), indicating that surviving primary tumor cells were dormant for an exceptionally long time, before the relapsing clone was able to expand. Whilst there was no evidence for reversion of the somatic *BRCA1* mutation, the emergent clone at recurrence lacked the *RB1* deletion present in the primary tumor (Fig. 2d), which may explain the relatively short duration of responses to chemotherapy at relapse, given the association between co-mutation of *BRCA1* and *RB1* with exceptional survival in HGSC[25]. In contrast, two patients (MAOC00944 and MAOC01893) maintained co-mutation of *BRCA1* and *RB1* at relapse, and were both alive at last follow up (Figs. 2a, c). Surprisingly, the metastatic brain tumor sample in MAOC00944 had acquired a different *RB1* deletion to the primary tumor (Fig. 2d), perhaps indicating that while *RB1*

inactivation is common and advantageous to HGSC growth, it is also associated with treatment vulnerability and exceptional outcomes.

For two cases, RNA sequencing was also performed on relapse samples. MAOC01893 underwent a switch from C2/immunoreactive to C1/mesenchymal molecular subtype, which was also associated with loss of chemokine amplification (*CXCL9*, *CXCL10*, *CXCL11*), between primary and relapse tumor samples (Fig. 2c). By contrast, patient MWMH00758 maintained the C2/immunoreactive subtype and chemokine amplification (*CXCL9*, *CXCL10*, *CXCL11*) from the primary tumor to two subsequent relapses (sampled at 6.4 and 8.6 years after diagnosis), indicating that this patient had a sustained immune response that along with chemotherapy and surgery augmented disease remission for more than 10 years.

1.6 *BRCA1*-altered tumors

Unsupervised clustering of primary tumor genomes based on their composition of mutational signatures stratified *BRCA1* altered tumors into three groups with different outcomes (Fig. 3, Extended Data Fig. 3b). Cluster SIG.7 had the highest proportion of long-term survivors (75%, 12/16), followed by SIG.6 (56%, 5/9) and SIG.5 (27%, 6/22). In order to identify key features of these groups, we performed a subset analysis of genomic and clinical characteristics, considering only the *BRCA1*-altered clusters. The key mutational signatures driving these clusters, in order of significance, were the double base substitution signature 2 (DBS2) in SIG.5 (P_{adj} = 0.0050), rearrangement signature Ovary_A in SIG.6 (P_{adj} = 0.0092), and single base substitution signature 40 (SBS40) in SIG.7 (P_{adj} = 0.0092; Supplementary Fig. 4).

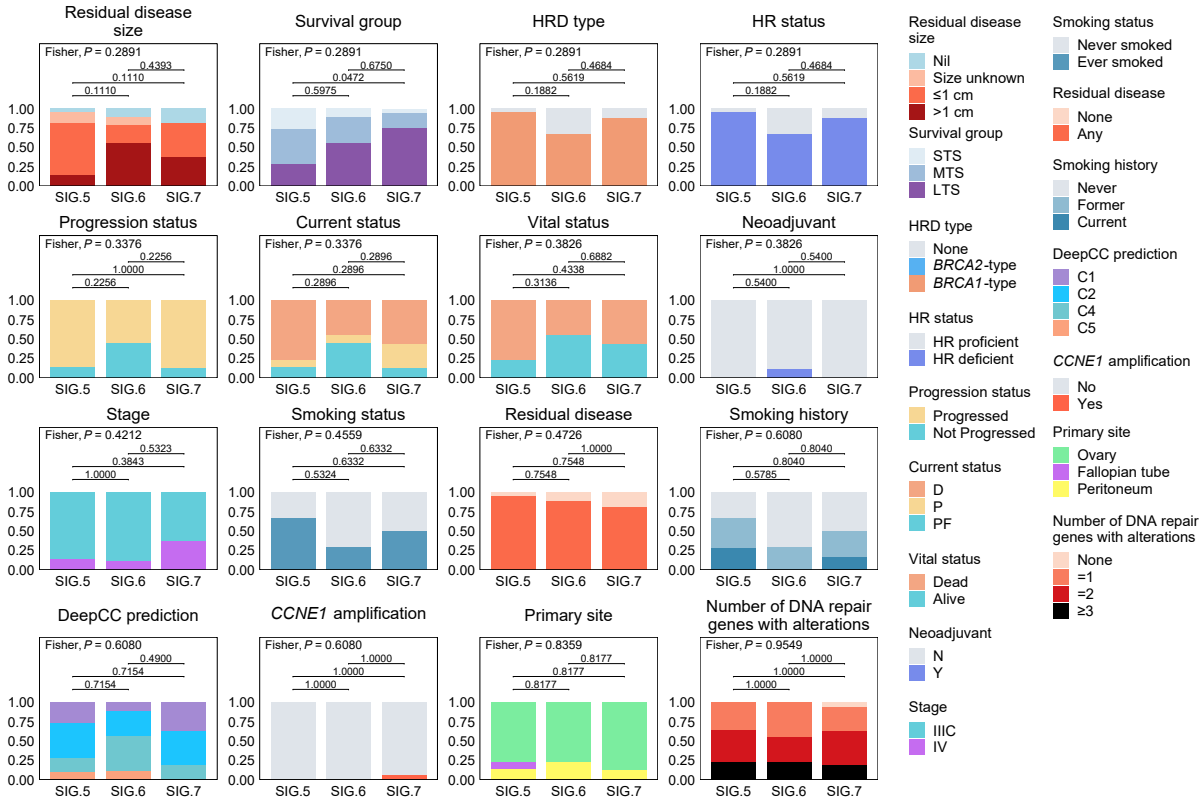


Supplementary Fig. 4 | Relative enrichment of mutational signatures in *BRCA1*-altered signature clusters. Boxplots show the relative proportion (y axis) of mutational signatures across *BRCA1*-altered mutational signature clusters (SIG.5 $n = 22$, SIG.6 $n = 9$, SIG.7 $n = 16$); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, whiskers show the smallest/largest values within 1.5 times the interquartile range, red triangles indicate the mean, and dotted lines join the means of each cluster to visualize the trend. Pairwise Mann-Whitney-Wilcoxon test adjusted *P* values (two-sided) and Kruskal-Wallis test adjusted *P* values are reported. Signatures are ordered left to right and top to bottom by their significance.

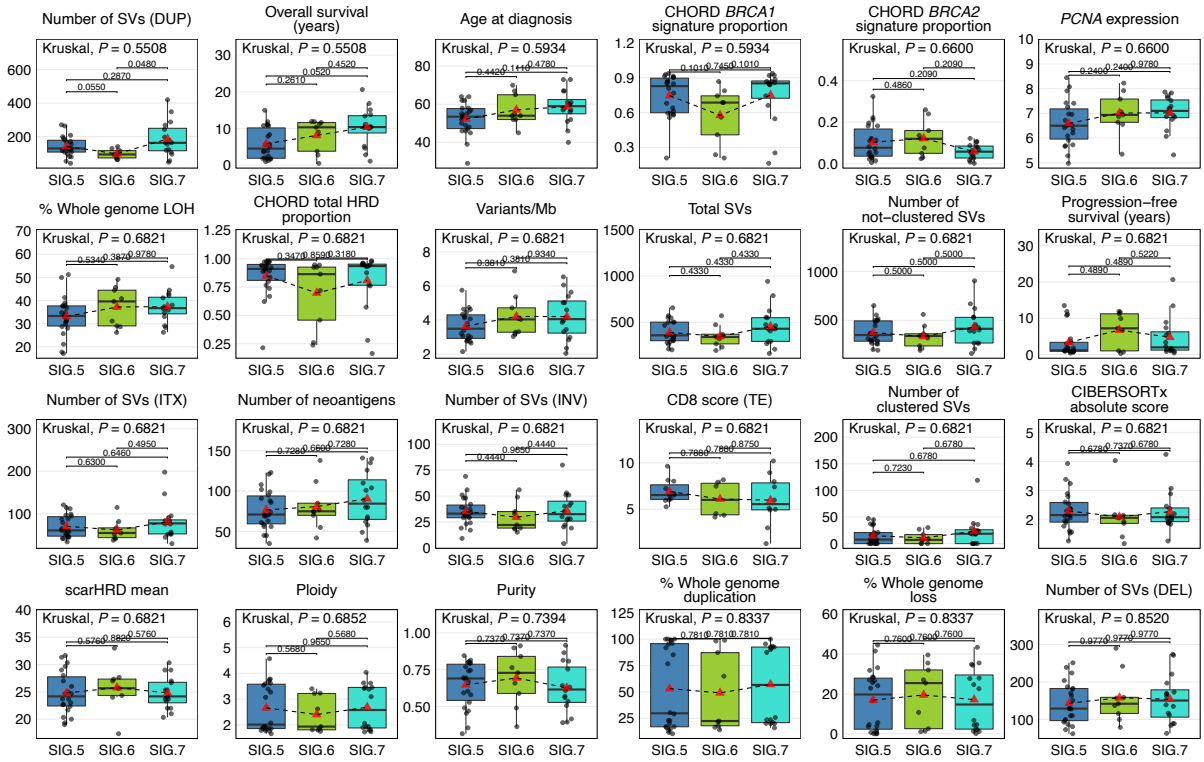
DBS2 is proposed to be associated with tobacco smoking and/or other mutagens[26], and is further explored in section 1.8 below. SBS40 is a relatively featureless signature seen in various cancer types that has not yet been linked to a plausible underlying cause[26]. Rearrangement signature Ovary_A is associated with *BRCA2* alterations[27], and its detection in the SIG.6 *BRCA1*-altered cluster indicates there may be combined *BRCA1* and *BRCA2* deficiency in tumors in this subgroup. Indeed, there are multiple tumors in SIG.6 with a mixture of both *BRCA1*-type and *BRCA2*-type HRD (CHORD scores, Fig. 3), and there is a higher prevalence of *BRCA2*-type non-clustered 1-100 kb deletions in SIG.6 compared to SIG.5 and SIG.7 (Extended Data Fig. 3d). It is plausible that tumors with combined *BRCA1* and *BRCA2*

loss of function may have greater sensitivity to DNA damaging chemotherapies, and this may explain why some HGSC patients with *BRCA1*-driven tumors have a better survival than others.

There were no significant differences in clinical features between *BRCA1*-altered clusters, although there was a trend towards bulkier residual disease (>1 cm) in cluster SIG.6 (P_{adj} = 0.2891; Supplementary Fig. 5) and a lower age of diagnosis in cluster SIG.5 (P_{adj} = 0.5934; Supplementary Fig. 6).



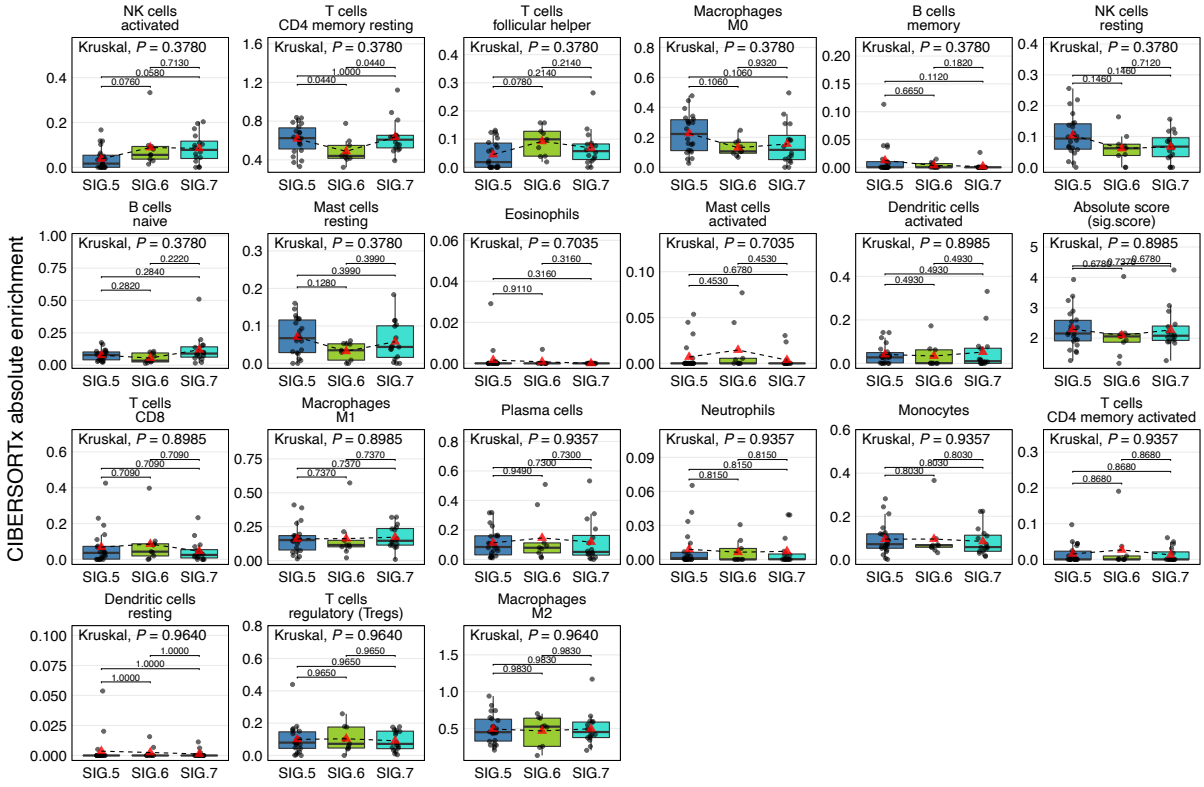
Supplementary Fig. 5 | Categorical clinical and molecular features of *BRCA1*-altered signature clusters. Proportion of patients with each feature across *BRCA1*-altered mutational signature clusters (SIG.5 n = 22, SIG.6 n = 9, SIG.7 n = 16). Fisher's exact test (two-sided) adjusted P values are reported, and features are ordered by significance. Features include homologous recombination (HR) status, homologous recombination deficiency (HRD) type, number of DNA repair pathway alterations, survival group (LTS, long-term survivor; MTS, moderate-term survivor; STS, short-term survivor), status at last follow-up (D, dead; P, progressed and alive; PF, progression-free and alive), self-reported smoking status, molecular subtype (C1, mesenchymal; C2, immunoreactive; C4, differentiated; C5, proliferative), and neoadjuvant treatment (Y, yes; N, no).



Supplementary Fig. 6 | Clinical and genomic features of *BRCA1*-altered signature clusters. Boxplots summarize numerical clinical and genomic features across *BRCA1*-altered mutational signature clusters (SIG.5 n = 22, SIG.6 n = 9, SIG.7 n = 16); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, whiskers show the smallest/largest values within 1.5 times the interquartile range, red triangles indicate the mean, and dotted lines join the means of each cluster to visualize the trend. Pairwise Mann-Whitney-Wilcoxon test adjusted P values (two-sided) and Kruskal-Wallis test adjusted P values are reported. Features are ordered by their significance and clusters are ordered by the proportion of long-term survivors. CD8 scores were measured by immunohistochemistry[25] and scored as density of CD8+ T cells (average cells/mm^2, y axis) in the tumor epithelium (TE). HRD, homologous recombination deficiency; DEL, deletions; DUP; duplications; SV, structural variants; Mb, megabase; ITX, interchromosomal translocations; LOH, loss-of-heterozygosity; INV, inversions.

Immune cell composition inferred from gene expression data[28] was compared between the *BRCA1*-altered clusters (Supplementary Fig. 7). While the majority of predicted immune cell types were relatively similar, the largest difference was observed in the activated natural killer (NK) cell signature, which was particularly elevated in clusters SIG.6 and SIG.7

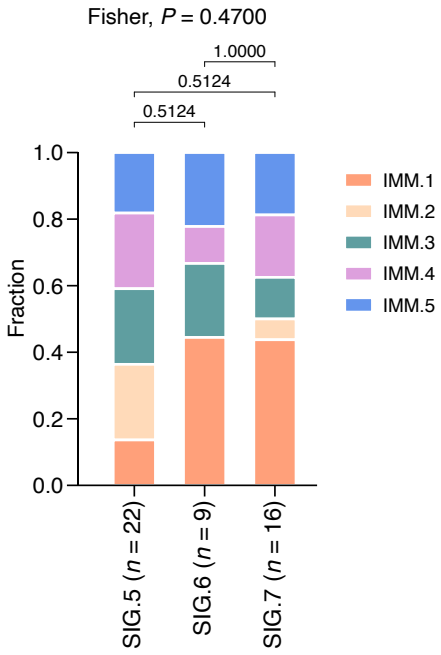
compared to SIG.5 (SIG.5 vs SIG.6 P_{adj} = 0.0760, SIG.5 vs SIG.7 P_{adj} = 0.0580). Although not significant, the resting NK cell signature showed the opposite trend, with an elevated score in cluster SIG.5 compared to SIG.6 and SIG.7 (SIG.5 vs SIG.6 P_{adj} = 0.1460, SIG.5 vs SIG.7 P_{adj} = 0.1460).



Supplementary Fig. 7 | Immune cell subsets inferred from gene expression data. Boxplots indicate the estimated abundance (y axis) of each immune cell type across *BRCA1*-altered mutational signature clusters (SIG.5 *n* = 22, SIG.6 *n* = 9, SIG.7 *n* = 16); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, whiskers show the smallest/largest values within 1.5 times the interquartile range, red triangles indicate the mean, and dotted lines join the means of each cluster to visualize the trend. Pairwise Mann-Whitney-Wilcoxon test adjusted *P* values (two-sided) and Kruskal-Wallis test adjusted *P* values are reported. Cell types are ordered left to right and top to bottom by their significance.

NK cell infiltration is associated with a favorable prognosis in various solid cancers[29], and a recent study found that while NK cells are relatively rare in HGSC, patients with a high density of tumor-infiltrating NK cells had a longer overall survival[30]. A higher activated NK cell signature was associated with longer survival in the overall cohort (hazard ratio (HR) =

0.67, 95% confidence interval (CI) = 0.44-1.04, $P = 0.0717$; Extended Data Fig. 7c) and could, at least partially, explain the longer overall survival observed in clusters SIG.6 and SIG.7 compared to SIG.5. Concordantly, immune cluster IMM.1, which is enriched with the activated NK cell signature (Fig. 5, Extended Data Fig. 8a), was the dominant immune cluster in both SIG.6 (44.4%, 4/9) and SIG.7 (43.8%, 7/16), and the least abundant immune cluster in SIG.5 tumors (13.6%, 3/22; $P = 0.4700$, Supplementary Fig. 8).



Supplementary Fig. 8 | Immune clusters in *BRCA1*-altered tumors. The fraction of tumors in each immune cluster IMM.1, IMM.2, IMM.3, IMM.4 or IMM.5, according to unsupervised clustering shown in Fig. 5. Samples are grouped according to their *BRCA1*-altered mutational signature clusters (SIG.5 $n = 22$, SIG.6 $n = 9$, SIG.7 $n = 16$). Differences in proportions between groups were assessed by Fisher's test (two-sided).

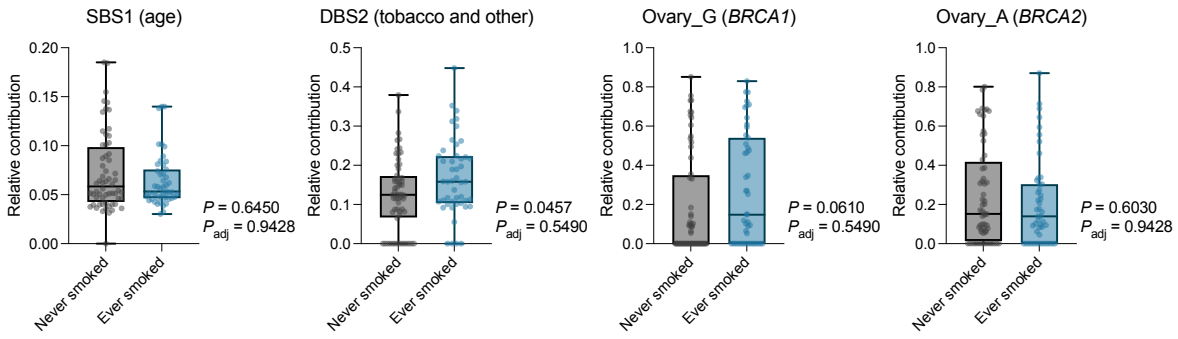
1.7 Smoking analysis

The poor survival *BRCA1* mutant SIG.5 subgroup was enriched for the double base substitution mutational signature 2 (DBS2; Fig. 3), which is thought to be associated with exposure to tobacco smoking and/or other mutagens such as acetaldehyde[26]. We sought evidence of other mutational signatures known to be associated with tobacco smoking (SBS4 and ID3) across the cohort of 126 primary tumor samples, however these signatures were present in only a few samples at a low level (mean relative contribution 0.005 and 0.026 respectively), and were

therefore excluded from the mutational signature fitting and clustering analyses (see Supplementary Methods).

Self-reported smoking history was available for 84.9% (107/126) of participants. Across the seven mutational signature subgroups, the poor survival *BRCA1*-mutant SIG.5 subgroup had the highest frequency of ever smokers (66.7%, 12/18), followed by SIG.1 (50%, 7/14) and SIG.7 (50%, 6/12; P_{adj} = 0.5092; Extended Data Fig. 4b). Across the survival groups, tobacco smoking was more common in moderate-term survivors (50%, 16/32), compared to short-term (38%, 13/34) and long-term survivors (30%, 18/60; Chi-square P = 0.1433; Extended Data Fig. 1b).

The relative contribution of all mutational signatures was compared between never smokers (n = 60) and ever smokers (n = 47), and the most predominant mutational signatures in smokers were DBS2 (P = 0.0457) and Ovary_G (*BRCA1* rearrangement signature; P = 0.0610), however these associations were not significant following multiple correction possibly due to the small number of cases (Supplementary Fig. 9, Supplementary Table 16). Altogether, there is evidence that the DBS2 mutational signature observed in our cohort may be associated with tobacco smoking.



Supplementary Fig. 9 | Association between pre-diagnosis smoking and mutational signatures. Box plots summarize the relative contribution (y-axis) of selected mutational signatures in never smokers (n = 60) and ever smokers (n = 47); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, and whiskers show the smallest and largest values. The proposed aetiology of each signature is shown in parentheses. SBS1 and Ovary_A were included in the figure as predominant mutational signatures in HGSC that are not associated with smoking status. Mann-Whitney test (two-tailed) was used to compare relative signature prevalence between groups, and both unadjusted and Benjamini-Hochberg adjusted P values (P_{adj}) are displayed. Results for all mutational signatures are shown in Supplementary Table 16.

In the landmark pan-cancer mutational signature analysis[26], SBS4, ID3 and DBS2 were detected at high levels in cancers caused by direct tobacco smoke exposure, for example lung and head and neck cancers. Given SBS4 and ID3 were not predominant in our cohort, it is not clear if the DBS2 mutations in our cohort are directly linked to tobacco smoking. Mutations consistent with DBS2 have been experimentally linked to acetaldehyde[31], a known carcinogen that is a constituent of cigarette smoke and a metabolite of alcohol. Tobacco and alcohol intake are correlated[32], so it is possible that smoking status in our cohort may be a proxy for alcohol intake, and that DBS2 could be associated with acetaldehyde exposure from drinking alcohol. Alcohol exposure information was not available for participants in this study at the time of writing, but the potential combined effects of acetaldehyde exposure and homologous recombination deficiency merits further investigation.

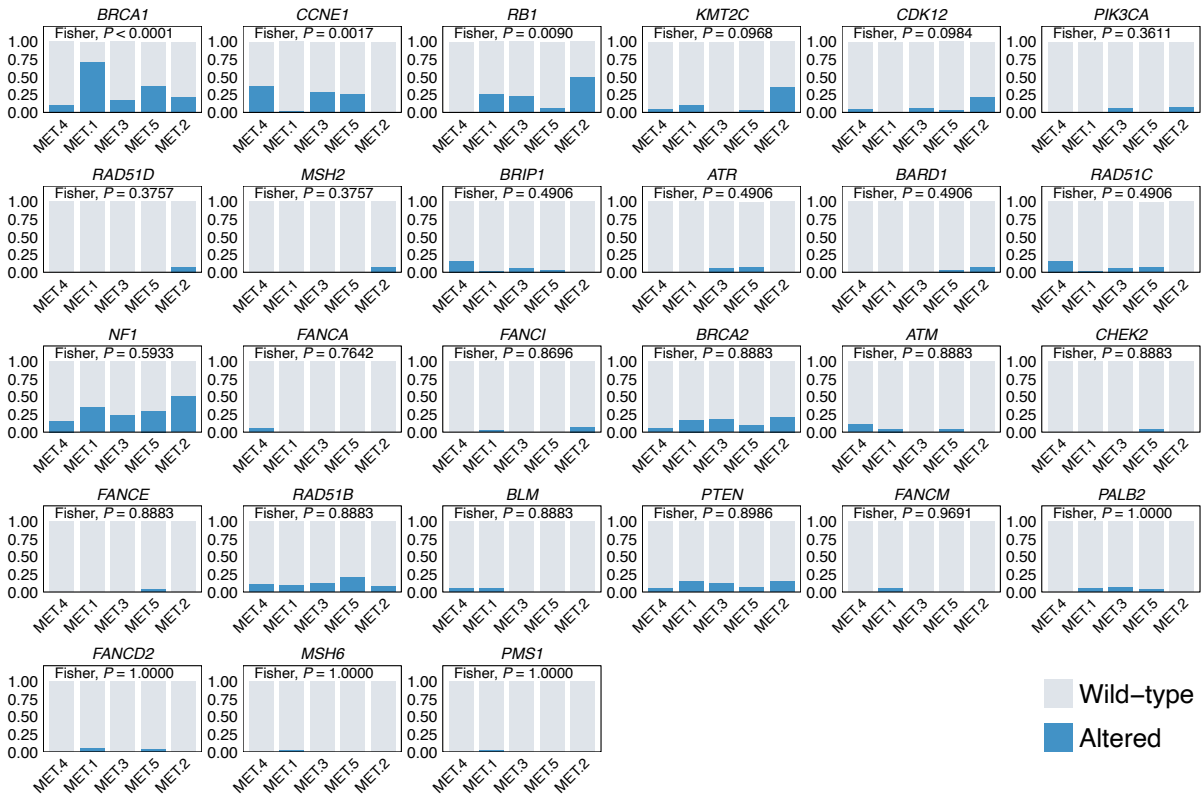
1.8 Methylation analysis

Gene inactivation by epigenetic silencing is a driver of HGSC[10]. To determine whether exceptional survival is associated with distinct patterns of tumor DNA methylation, we examined genome-wide differential DNA methylation between each survival group (Supplementary Data 6). Filters were applied to select CpG probes that negatively correlated with gene expression across the cohort of 126 primary tumors. For the complete differential methylation analysis workflow and filtering steps see Supplementary Methods.

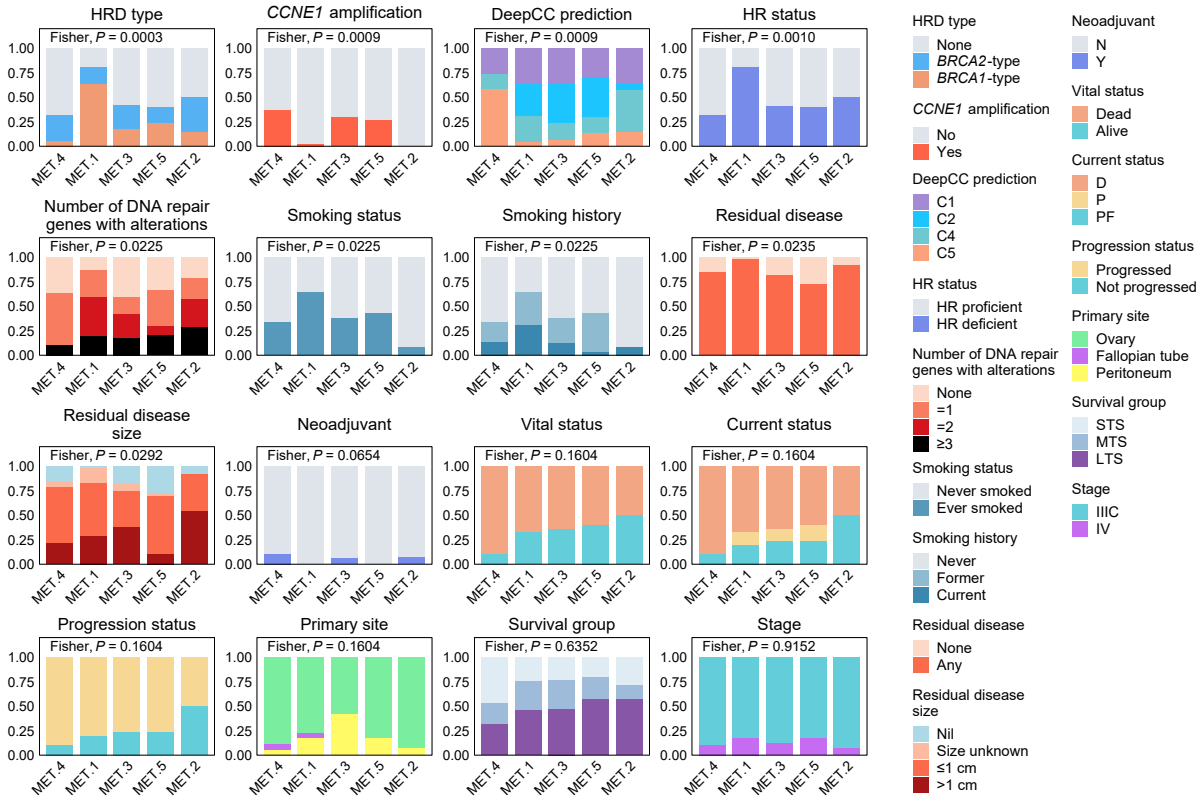
There were 34 probes, associated with 25 protein coding genes, identified as differentially methylated between long-term survivors (n = 60) and moderate-term survivors (n = 32). Among these 25 genes, four were identified as having reduced expression in long-term survivors compared to moderate-term survivors: *ABT1*, *CPE*, *PAIP2B* and *PRSS23* (DESeq2 $P_{adj} < 0.1$, $\log_2$ fold change < 0; Supplementary Data 4). Comparing long-term survivors and short-term survivors (n = 34), we identified 43 differentially methylated probes, associated with 29 genes. Three of these genes (*CPE*, *PAIP2B*, *PRSS23*) also had significantly lower expression in long-term survivors compared to short-term survivors (DESeq2 $P_{adj} < 0.1$, $\log_2$ fold change < 0; Supplementary Data 4). Despite being clinically distinct, we did not identify differentially methylated probes between moderate and short-term survivors. This is comparable with our previous analysis[14], in which we did not identify significantly differentially methylated loci between primary sensitive and primary resistant/refractory HGSC.

Of the 37 genes associated with methylated loci in long-term survivors, three have been causally implicated in human cancer[19]: *EPHA7*, *FGFR1*, and *SLC34A2*. *SLC34A2* (NaPi2b) is of particular interest, as it has been implicated as a tumor suppressor gene in several cancer types[33,34], is a potential new therapeutic target in ovarian cancer[35-37], and decreased expression is associated with enhanced chemosensitivity in breast cancer[38]. *EPHA7* mutations are associated with enhanced anti-tumor immunity and better clinical outcomes in patients treated with immune checkpoint inhibitors, in multiple cancer types[39]. Additional studies will be required to determine which of these candidates are epigenetically modified in HGSC, and to explore their association with clinical outcomes.

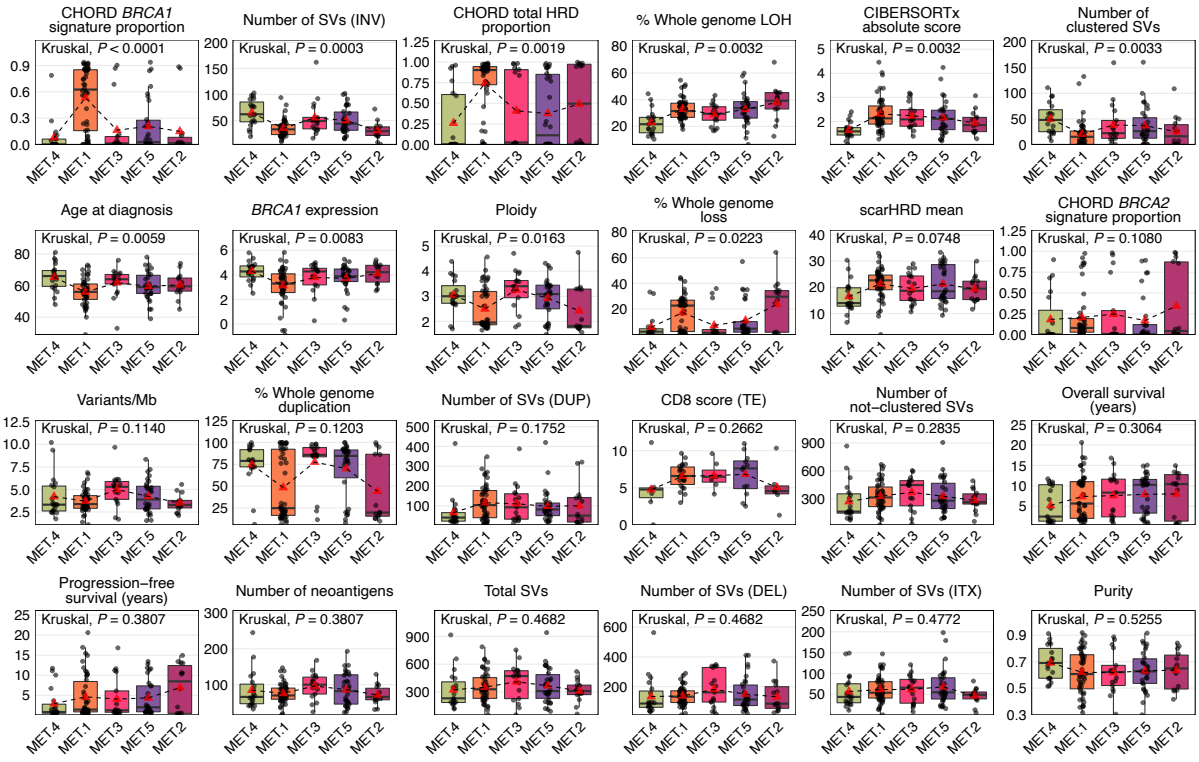
We performed consensus clustering of the 1% most variable CpG probes (number of probes = 3,645) across all 126 primary tumors, and identified five clusters with moderate to weak correlations with progression-free (P = 0.1949) and overall survival (P = 0.0587, Extended Data Fig. 6). Cluster MET.2 had the longest overall survival (median OS 11.9 years), followed by MET.5 (median OS 10.4 years), MET.3 (median OS 8.4 years), MET.1 (median OS 5.7 years), and MET.4 (median OS 2.5 years). Across all genomic and clinical characteristics considered in the study, we identified *BRCA1* alterations as the key difference between methylation clusters, with a particularly high proportion of *BRCA1*-altered tumors in cluster MET.1 (71.7%, n = 33/46; $P_{adj} < 0.0001$, Supplementary Fig. 10). Concordantly, MET.1 had a high proportion of *BRCA1*-type HRD (P_{adj} = 0.0003, Supplementary Fig. 11) and the lowest *BRCA1* expression of all the clusters (P_{adj} = 0.0083, Supplementary Fig. 12). This finding is consistent with previous studies that identified an association between *BRCA1* inactivation and distinct genome-wide methylation patterns in breast cancer[40,41], suggesting that *BRCA1* deficiency may lead to global DNA methylation changes.



Supplementary Fig. 10 | Gene alterations across DNA methylation clusters. Proportion of patients affected by gene alterations per methylation cluster. Genes are ordered by significance using Fisher's exact test (two-sided) and clusters are ordered by the proportion of long-term survivors. MUT, mutated; WT, wild-type. The Fisher's test *P* values displayed are Benjamini-Hochberg adjusted *P* values.



Supplementary Fig. 11 | Proportion of genomic and clinical features by DNA methylation cluster. Features are ordered by significance using Fisher's exact test (two-sided) and the clusters are arranged by the proportion of long-term survivors. *P* values are Benjamini-Hochberg adjusted. Features include homologous recombination (HR) status, homologous recombination deficiency (HRD) type, number of DNA repair pathway alterations, self-reported smoking status, survival group (LTS, long-term survivor; MTS, moderate-term survivor; STS, short-term survivor), status at last follow-up (D, dead; P, progressed and alive; PF, progression-free and alive), molecular subtype (C1, mesenchymal; C2, immunoreactive; C4, differentiated; C5, proliferative), and neoadjuvant treatment (Y, yes; N, no).



Supplementary Fig. 12 | Clinical and genomic features of DNA methylation clusters. Boxplots summarize numerical clinical and genomic features across methylation clusters (MET.4 n = 19, MET.1 n = 46, MET.3 n = 17, MET.5 n = 30, MET.2 n = 14); points represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, whiskers show the smallest/largest values within 1.5 times the interquartile range, red triangles indicate the mean, and dotted lines join the means of each cluster to visualize the trend. Pairwise Mann-Whitney-Wilcoxon test adjusted P values (two-sided) and Kruskal-Wallis test adjusted P values are reported. Features are ordered by their significance and clusters are ordered by the proportion of long-term survivors from left to right. HRD, homologous recombination deficiency; DEL, deletions; DUP; duplications; SV, structural variants; Mb, megabase; ITX, interchromosomal translocations; LOH, loss-of-heterozygosity; INV, inversions. CD8 scores were available for n = 54 primary tumors as measured by immunohistochemistry[25] and scored as density of CD8+ T cells (average cells/mm^2, y axis) in the tumor epithelium (TE).

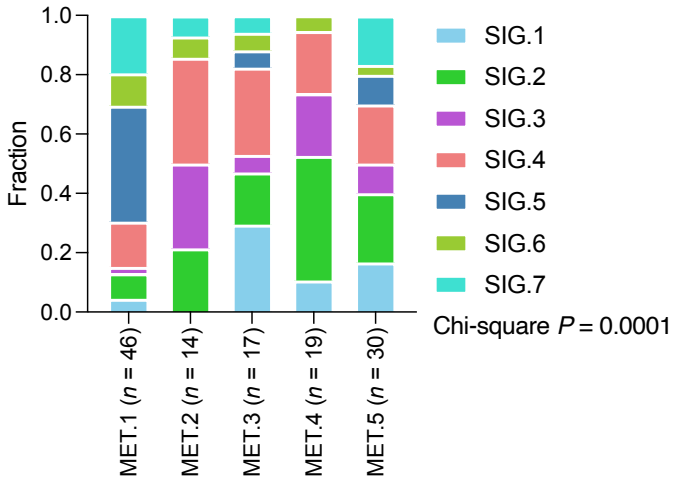
Given the frequency of somatic methylation of the *BRCA1* 5' UTR CpG island North Shore in HGSC[10], we wondered if the methylation clusters could be driven by differential methylation of the *BRCA1* CpG sites. However, of the 3,645 probes in the cluster analysis, only

two were *BRCA1* promoter CpG sites (cg25288140, cg15065591), neither of which were used in the classification of *BRCA1* promoter methylation.

The MET.2 subset had the best prognosis, and was enriched for alterations in *RB1* (P_{adj} = 0.0090), *KMT2C* (P_{adj} = 0.0968), and *CDK12* (P_{adj} = 0.0984, Supplementary Fig. 10). MET.4, the methylation cluster associated with the poorest survival, was characterised by the lowest CIBERSORTx absolute score (P_{adj} = 0.0032, Supplementary Fig. 12), the most frequent amplification of *CCNE1* (37%, 7/19; P_{adj} = 0.0009), and enrichment for the C5/proliferative molecular subtype (P_{adj} = 0.0009, Supplementary Fig. 11), which is known to be immune depleted and associated with unfavorable outcomes[42,43]. MET.4 likely corresponds with the methylation cluster associated with the poorest survival in TCGA[10], DNA Methylation Subtype 2 or "MC2", which was also significantly enriched for the C5/proliferative molecular subtype.

CCNE1 amplification status appeared to differentiate the two main branches of the methylation dendrogram (Extended Data Fig. 6a), as the three clusters constituting the right branch (MET.3, MET.4, and MET.5) all had a high proportion of *CCNE1* amplification (27% to 37%) compared to the clusters on the left branch (MET.1 and MET.2), which had a *CCNE1* amplification rate of 2% (1/46) and 0% (0/14) respectively (P_{adj} = 0.0009; Supplementary Fig. 11). In addition to lacking *CCNE1* amplification, MET.1 and MET.2 also had a high proportion of tumors with multiple DNA repair alterations (P_{adj} = 0.0225; Supplementary Fig. 11), and a high proportion of whole genome loss (P_{adj} = 0.0223; Supplementary Fig. 12).

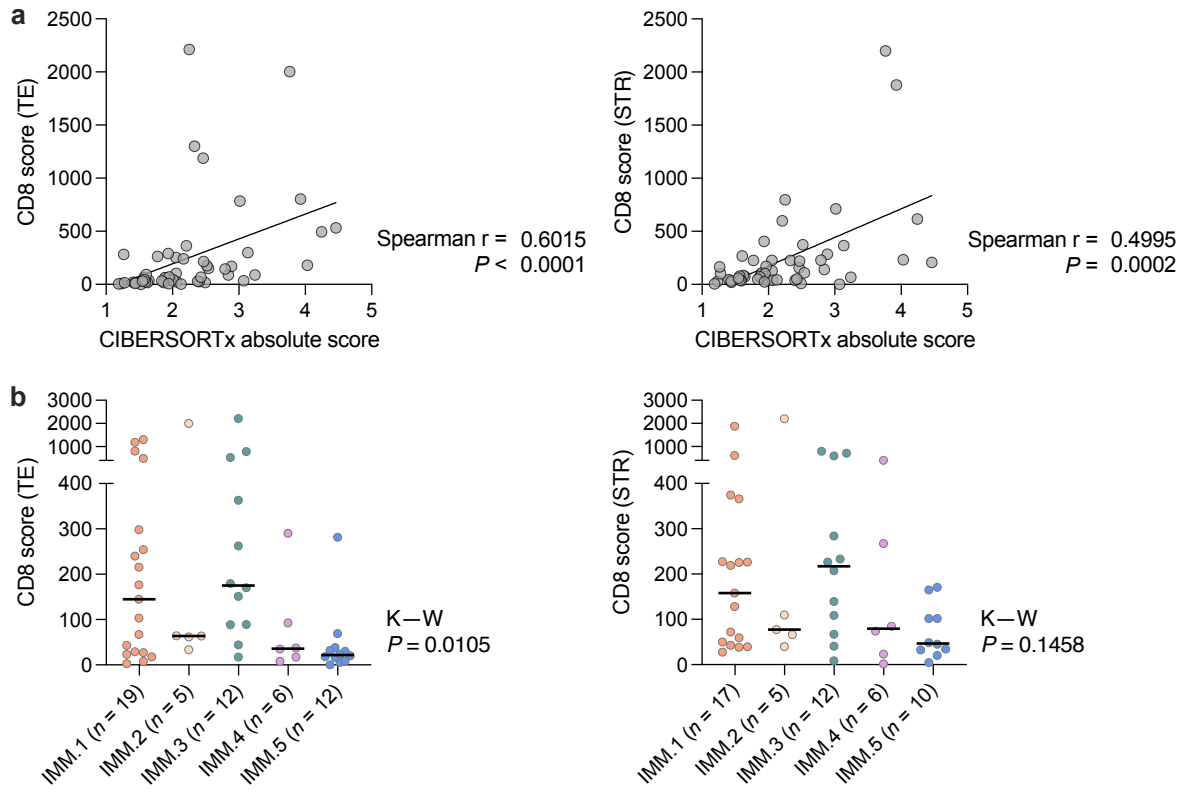
Despite their similarities, patients in MET.1 and MET.2 had vastly different survival outcomes (median OS 5.7 vs. 11.9 years), so we contrasted their clinical and molecular features, finding that MET.1 had the largest proportion of smokers (64%, 23/36) and MET.2 the lowest (8.3%, 1/12; P_{adj} = 0.0225, Supplementary Fig. 11), indicating a potential link between pre-diagnosis smoking exposure and differential methylation patterns. Exposure to cigarette smoke can result in DNA methylation changes detected in peripheral blood[44] and lung cancer samples[45], and some studies have linked tobacco exposure to methylation changes in distal sites such as bladder cancer[46] and adipose tissue[47], however little is currently known about smoking associated DNA methylation alterations in ovarian cancer. Similar to the smoking-associated cluster SIG.5, the MET.1 subgroup also had a younger age of diagnosis (P_{adj} = 0.0059, Supplementary Fig. 12). Indeed, the MET.1 cluster is enriched for mutational signature cluster SIG.5 tumors (Supplementary Fig. 13), indicating that the DNA methylation clustering analysis has also identified the poor survival subset of *BRCA1*-altered tumors associated with smoking and a younger age of diagnosis.



Supplementary Fig. 13 | Mutational signature clusters versus DNA methylation clusters. The proportion of tumors in each mutational signature cluster according to consensus clustering shown in Fig. 3a. Samples are grouped according to their DNA methylation clusters. Differences in proportions between groups were assessed by Chi-square test.

1.9 Immune cell estimation versus immunohistochemical staining

To evaluate the presence of immune cells in the tumor microenvironment, the CIBERSORTx[28] immune cell deconvolution method was used to estimate the relative abundance of 22 immune cell types. To verify the accuracy of this method, we compared the CIBERSORTx absolute score (an aggregate gene set score associated with overall immune activity) to the CD8+ T-cell counts, which were previously quantified by immunohistochemistry and image analysis[25] in a subset of tumors ($n = 54$ primary tumors). The CIBERSORTx absolute score showed a positive correlation with the density of CD8+ T cells, both in the tumor epithelium ($P < 0.0001$) and the stromal compartments ($P = 0.0002$, Supplementary Fig. 14a).

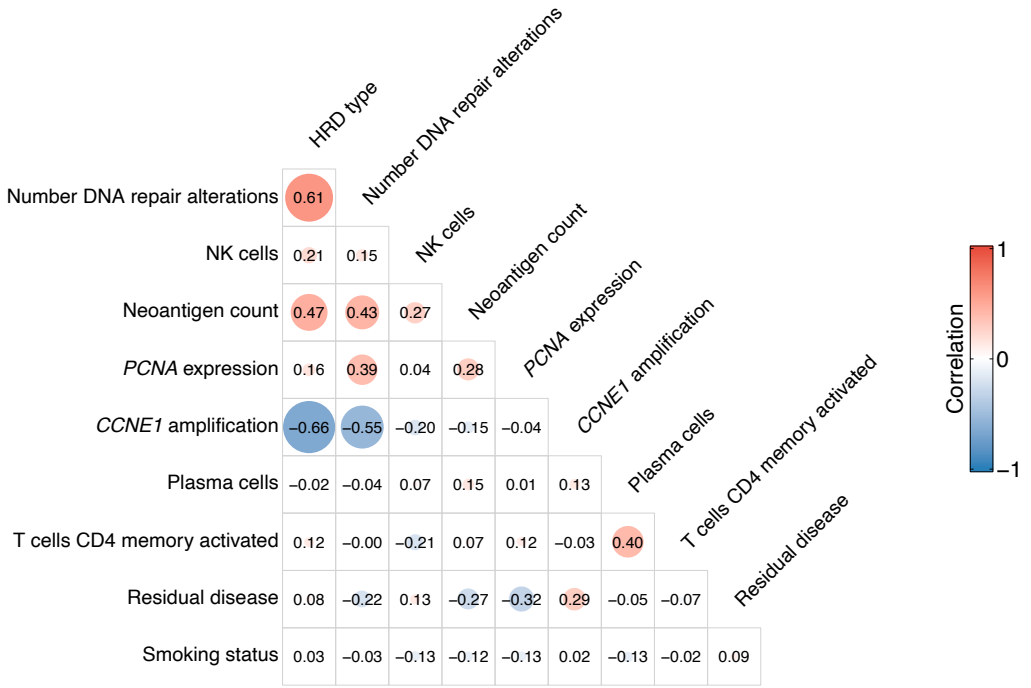


Supplementary Fig. 14 | Comparison of RNA-seq based immune cell estimation and immunohistochemical staining. a, Scatter plots show gene expression-based immune cell estimation scores (CIBERSORTx absolute scores, x axis) plotted against the density of CD8+ T cells (average cells/mm^2, y axis) as measured by immunohistochemistry staining in the tumor epithelium (TE; left) and stroma (STR; right). CD8+ T cell scores were available in the TE for n = 54 primary tumors, and a subset of n = 50 had CD8+ T cell counts in the STR. Spearman correlation P values reported (two-tailed). **b**, Density (in cells/mm^2) of CD8+ lymphocytes in TE and STR, grouped by the immune cluster category. Black lines indicated the median score; Kruskal-Wallis (K—W) test P values are reported with no adjustments for multiple comparisons.

Unsupervised clustering of primary tumors based on the estimated relative abundance of the most common immune cell types, stratified tumors into five groups with different outcomes (Fig. 5). The two immune clusters associated with the best prognosis, IMM.1 and IMM.3, were found to have higher densities of CD8+ T cells in both the tumor epithelium (P = 0.0105) and stromal compartments (P = 0.1458) compared to the other immune clusters (Supplementary Fig. 14b).

1.10 Feature independence

Some features associated with survival by univariate analysis were moderately correlated, including the number of DNA repair gene alterations and neoantigen count (0.43), and HRD type (*BRCA1*-type, *BRCA2*-type, none) and *CCNE1* amplification (-0.66), however most features were only weakly associated with each other (Supplementary Fig. 15). Features significantly associated with overall survival by univariate analysis (Fig. 6b) were also tested in a multivariable Cox regression model. Of the seven features entered in the model, four independent predictors remained: *PCNA* expression, plasma cells, HRD type and residual disease (Supplementary Tables 13 and 14).

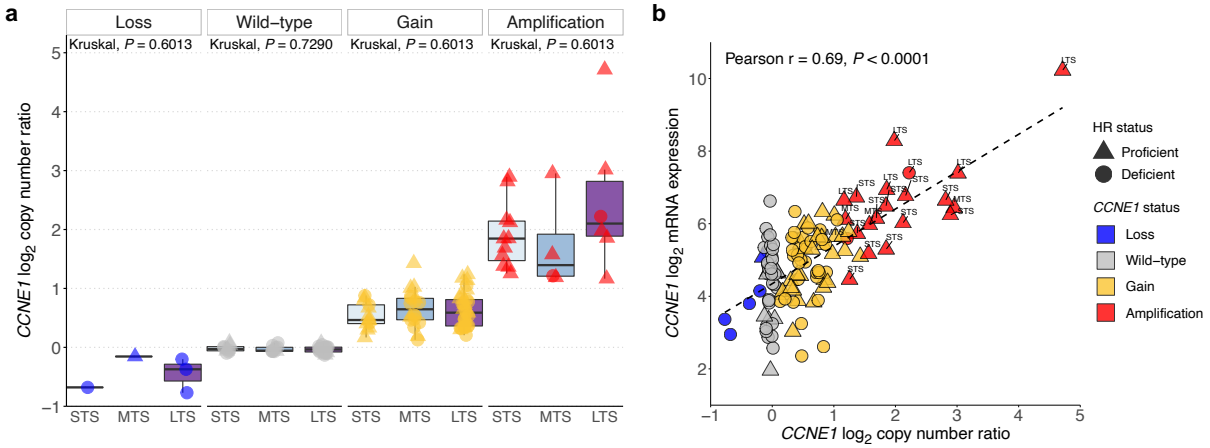


Supplementary Fig. 15 | Correlation of prognostic features. Correlation plot showing the pair-wise associations between individual prognostic factors. HRD, homologous recombination deficiency; NK, natural killer.

1.11 Long-term survivors with *CCNE1* amplification

We noted that 6 of the 60 long-term survivors had amplification of *CCNE1* in their primary tumor samples (Fig. 1b). This was an unexpected finding since focal *CCNE1* amplification is an established poor prognosis feature of HGSC, associated with primary platinum resistance and shorter patient survival[14,48]. Among long-term survivors with *CCNE1* amplification, 5 of

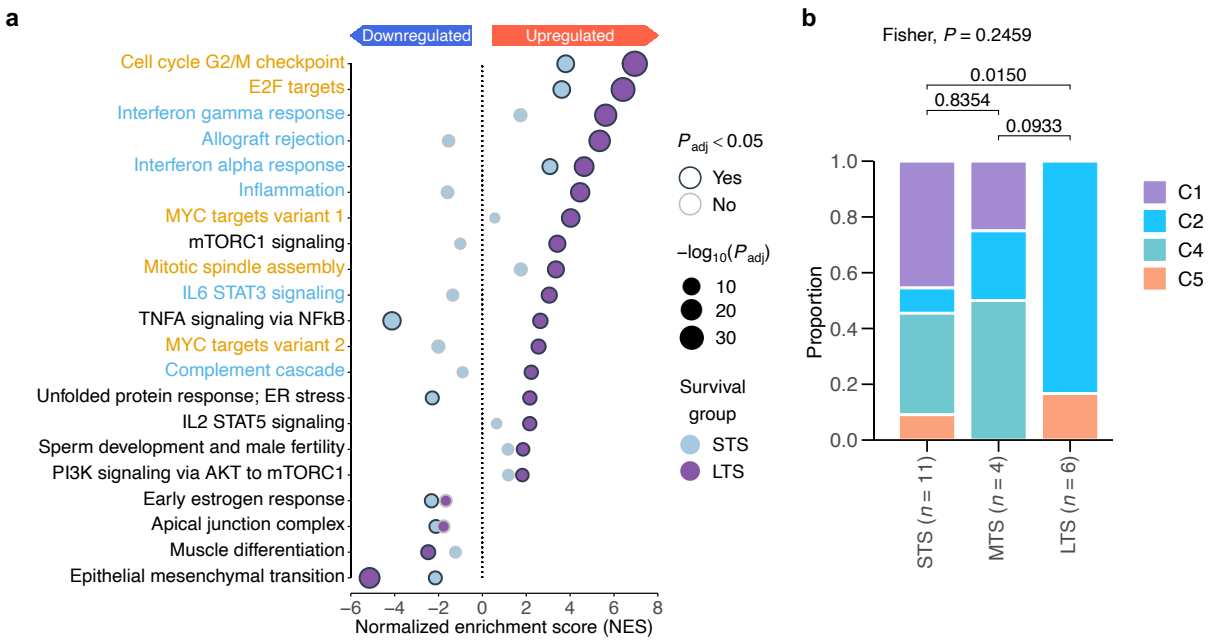
the 6 tumors had low CHORD scores consistent with being homologous recombination proficient, with one exception (MMAY00621) showing a clonal somatic *BRCA1* structural variant associated with an elevated *BRCA1*-type CHORD score. All *CCNE1* amplifications, regardless of survival group, were classified as clonal events (100%, 21/21; Supplementary Table 17). The distribution of *CCNE1* copy number in tumors with *CCNE1* amplification was equivalent between survival groups (P_{adj} = 0.6013, Supplementary Fig. 16a), and there was a strong correlation between *CCNE1* copy number and expression (Pearson r = 0.69, $P < 0.0001$; Supplementary Fig. 16b).



Supplementary Fig. 16 | *CCNE1* copy number and expression. a, Boxplots show *CCNE1* $\log_2$ copy number ratio (y-axis) in primary tumor samples, grouped by *CCNE1* copy number status and survival groups (n = 126); symbols represent each sample, boxes show the interquartile range (25-75th percentiles), central lines indicate the median, and whiskers show the smallest/largest values within 1.5 times the interquartile range. Kruskal-Wallis adjusted P values are reported. **b**, *CCNE1* $\log_2$ mRNA expression (y-axis) plotted against *CCNE1* $\log_2$ copy number ratio (x-axis) in primary tumor samples (n = 126; Pearson's correlation test, two-sided). Tumor samples with *CCNE1* amplification are labelled by survival group. In **a** and **b**, symbols represent each tumor sample, with shapes indicating homologous recombination (HR) status (triangle, proficient; circle, deficient) and colors showing the *CCNE1* copy number status. LTS, long-term survivor; MTS, moderate-term survivor; STS, short-term survivor.

To explore potential differences in *CCNE1* amplified tumors between survival groups, we contrasted transcriptomes of *CCNE1* amplified tumors in each survival group (short-term survivors n = 11, moderate-term survivors n = 4, long-term survivors n = 6) with a reference group of tumors with no *CCNE1* amplification or loss that had no homologous recombination

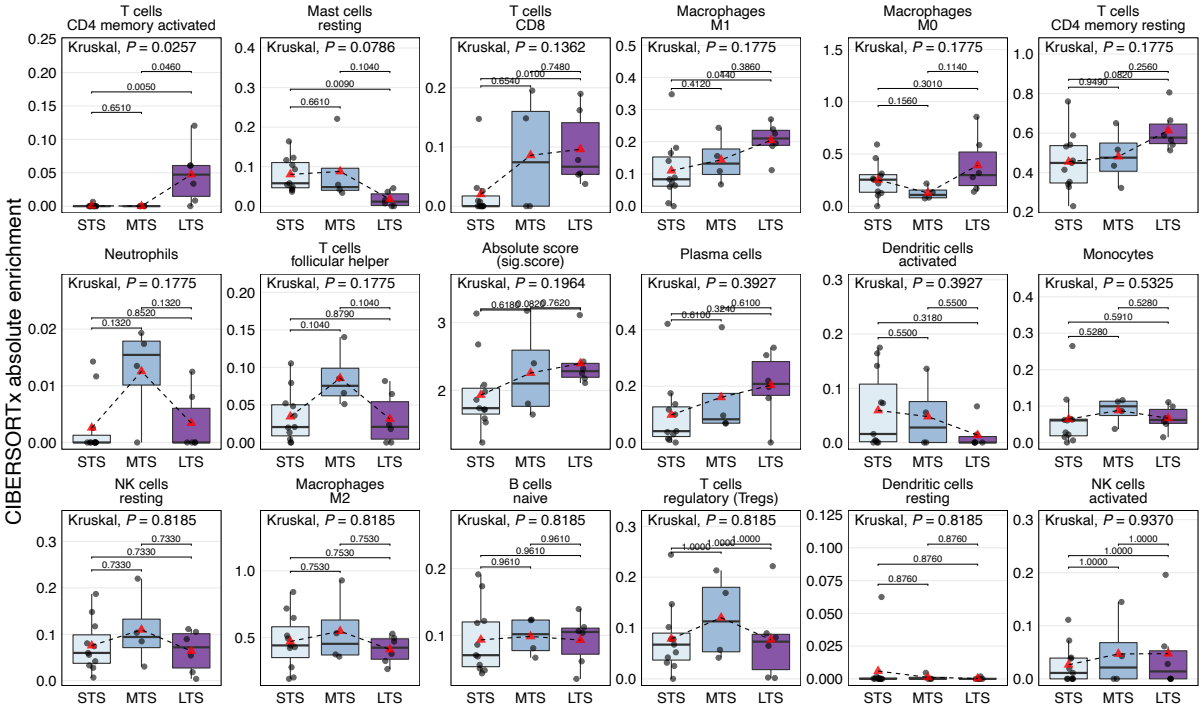
gene alterations and were classified as homologous recombination proficient (n = 21; Supplementary Data 7). Gene set enrichment analysis using the MSigDB Hallmark gene sets[49] found that *CCNE1* amplified tumors in long-term survivors were characterized by transcriptional activation of proliferation and immune-related pathways, and downregulation of epithelial mesenchymal transition (FGSEA $P_{adj} < 0.05$, Supplementary Fig. 17a). Consistent with this observation, tumors with *CCNE1* amplification in long-term survivors were predominantly the C2/immunoreactive[10,42] molecular subtype (83.3%, 5/6); a significant enrichment compared to the short-term survivors (9.09%, 1/11; $P_{adj} = 0.0150$, Supplementary Fig. 17b). Molecular subtypes in *CCNE1* amplified tumors in short-term survivors were mixed; although the C1/mesenchymal subtype, characterized by a desmoplastic stroma and poor survival[43], was most common in this group (45.5%, 5/11).



Supplementary Fig. 17 | Transcriptomic features of tumors with *CCNE1* amplification. a, MSigDB Hallmark gene sets enriched in long-term survivors (LTS, n = 6) or short-term survivors (STS, n = 11) with *CCNE1* amplification compared to a reference group of homologous recombination proficient tumors with no *CCNE1* amplification or loss and no homologous recombination gene alterations (n = 21). The normalized enrichment score (NES, x-axis) generated by FGSEA indicates the strength and direction of pathway activity (downregulated or upregulated) compared to the reference. P values (two-sided) were calculated using the FGSEA default Monte Carlo method; bubble size corresponds to the negative $\log_{10}$ Benjamini-Hochberg adjusted P value (P_{adj}) and bubble border color indicates significant enrichment (black = $P_{adj} < 0.05$, grey = $P_{adj} \geq 0.05$). Colored text indicates gene sets

associated with proliferation (gold) and immune activity (blue). Complete FGSEA results are listed in Supplementary Data 7. **b**, The proportion of tumors with *CCNE1* amplification classified as either C1/mesenchymal, C2/immunoreactive, C4/differentiated or C5/proliferative molecular subtype by survival group. Pairwise Mann-Whitney-Wilcoxon test adjusted *P* values and Fisher's test adjusted *P* values are reported (both two-sided). MTS, moderate-term survivor.

To further explore the enrichment of immune signatures, we performed gene expression-based immune cell estimation (CIBERSORTx) in the 21 tumors with *CCNE1* amplification (STS $n = 11$, MTS $n = 4$, LTS $n = 6$), finding significant enrichment of activated CD4 memory T cells ($P_{adj} = 0.0050$), CD8 T cells ($P_{adj} = 0.0100$) and M1 macrophages ($P_{adj} = 0.0440$) in long-term survivors compared to short-term survivors (Supplementary Fig. 18). This suggests that some patients, despite having features associated with primary treatment failure (i.e. homologous recombination intact tumors with *CCNE1* amplification), display features of an activated immune response and can attain long-term survival.



Supplementary Fig. 18 | Immune cell subsets in *CCNE1* amplified tumors. Boxplots indicate the estimated abundance (y axis) of immune cell types in *CCNE1* amplified tumors across survival groups (STS, short-term survivor $n = 11$; MTS, moderate-term survivors $n = 4$; LTS, long-term survivors $n = 6$); points represent each sample, boxes show the interquartile

range (25-75th percentiles), central lines indicate the median, whiskers show the smallest/largest values within 1.5 times the interquartile range, red triangles indicate the mean, and dotted lines join the means of each cluster to visualize the trend. Pair-wise Mann-Whitney-Wilcoxon test adjusted *P* values (two-sided) and Kruskal-Wallis test adjusted *P* values are reported. Cell types are ordered left to right and top to bottom by their significance, and only include cell types in which ≥5 samples had CIBERSORTx absolute enrichment scores >0.

We investigated other molecular features of *CCNE1* amplified tumors, finding no significant differences in mutational signature contributions or the number of genomic alterations between survival groups. Among long-term survivors with *CCNE1* amplification, one patient (MWMH00758) had amplification of the 4q13 chemokine gene cassette (containing *CXCL9*, *CXCL10* and *CXCL11*), and two other patents (AOCS-124, MMAY00010) had amplification of the cytokine gene *IFNG* (Fig. 1b). Although the sample sizes are small, we noted that two-thirds of *CCNE1* amplified long-term survivors (n = 4/6) also had *RAD51B* structural variants compared to 18% of short-term survivors (n = 2/11, Fig. 1b). There were no significant differences in clinical features between survival groups among cases with *CCNE1* amplification (age at diagnosis, primary site, grade, stage, residual disease, neoadjuvant therapy, smoking status), aside from features associated with survival group selection (current status, progression-free and overall survival; data not shown).

In summary, features associated with long-term survival, such as highly proliferative tumors or an active immune response, may override *CCNE1* driven resistance to primary chemotherapy.

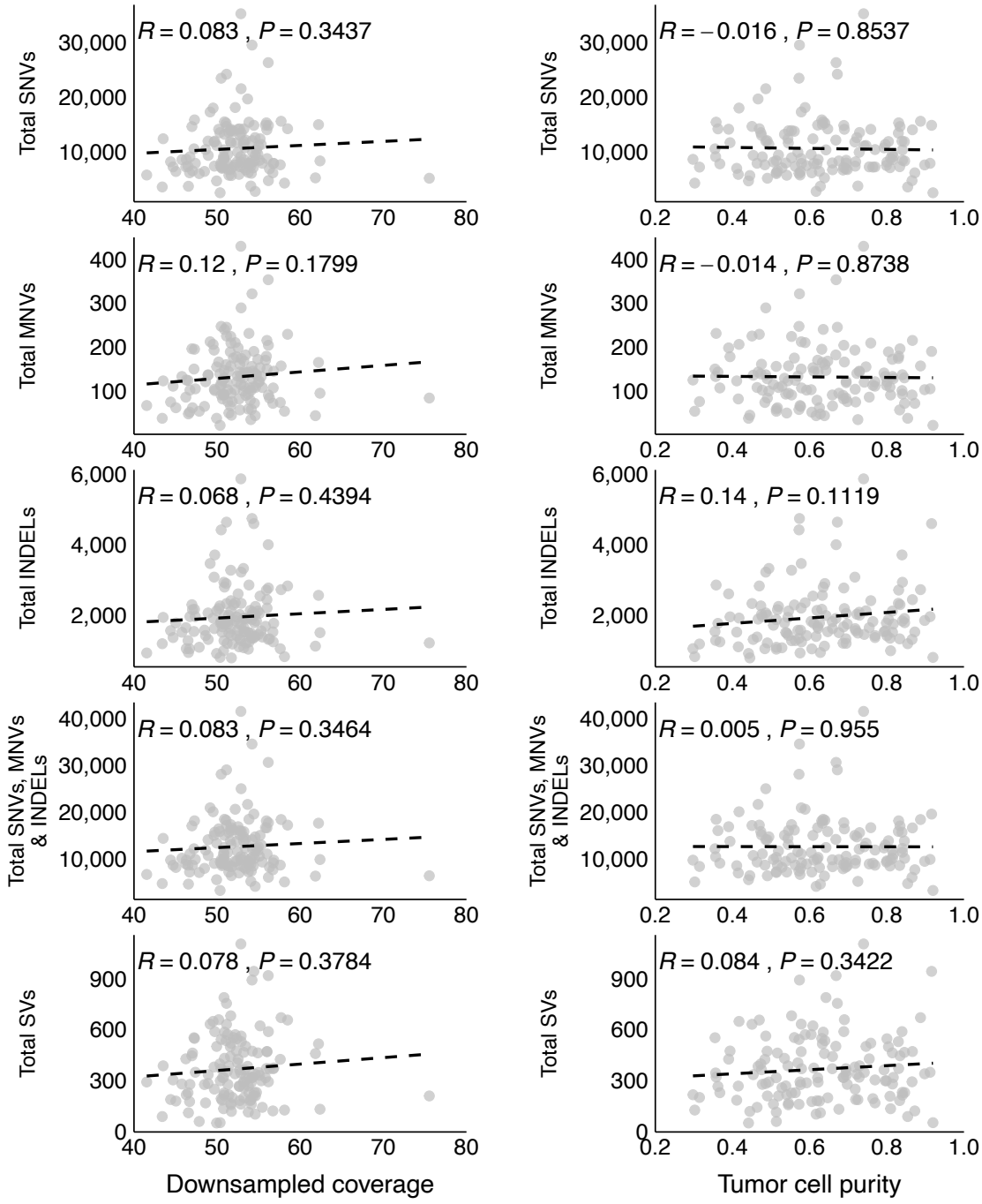
2. Supplementary Methods

2.1 Sample quality control

The median estimated tumor cell purity using FACETS[50] was 0.63 (range: 0.30 to 0.92; Supplementary Table 3). Following quality control checks, two tumor samples in the short-term survivor category were flagged as outliers with a high somatic mutation rate (>20 mutations/Mb). One due to cross sample contamination and the second due to a previously described germline mutation in the mismatch repair gene *PMS2*[14]. The two samples AOCS-076 and AOCS-166 were therefore excluded from further analyses.

2.2 Mutation burden and downsampling

To compare mutation burden across samples with variable coverage, higher coverage tumor BAMs were downsampled such that the median coverage between the previously sequenced samples[14] (ICGC cohort) and the recently sequenced samples (MOCOG cohort) was balanced. Downsampling was performed using Picard DownsampleSam (v2.17.3), which applies a downsampling algorithm to BAMs to retain a random subset of reads at a predetermined fraction. ICGC tumors had a median coverage of 52.15x and the MOCOG tumors had a median coverage of 77.81x. Thus, a downsampling fraction of 0.67 was used for the MOCOG BAMs to achieve equal median coverage with the ICGC samples. Tumor sample coverage before and after downsampling is listed in Supplementary Table 3, and following downsampling the number of called SNVs, MNVs, indels and SVs were independent of coverage (Supplementary Fig. 19).



Supplementary Fig. 19 | Correlation of coverage and tumor cell purity with observed somatic alterations. Scatter graphs show coverage (x-axis, left) and tumor cell purity (x-axis, right) plotted against the number of mutations (y-axis) detected in each tumor sample ($n = 131$), based on the down-sampled tumor genomes (Spearman correlation test, two-sided). Mutation types include: SNVs, single nucleotide variants; MNVs, multi-nucleotide variants; INDELs, small-scale insertions and deletions; SVs, structural variants.

2.3 Somatic base substitution and indel annotation

Blacklisted Duke and DAC mappability regions were annotated along with repeatmasker elements from the repeatmasker database using BCFtools[51] (v1.9). Strand specific read counts were added to the VCF using bam-readcount (v0.8.0 with -w 0 --min-mapping-quality 10 --min-base-quality 10 --max-count 100,000,000). Variants were annotated using the Ensembl Variant Effect Predictor[52] (VEP v92.4).

2.4 Structural variants (SVs)

SV annotation: SVs were annotated for genes using the R package csaw (v1.20.0) against the GRCh37.87 annotation release using the parameters (dist=5000, promoter=c(3000, 1000), ignore.strand=TRUE). Breakpoints were individually annotated to overlapping genes. A gene with any overlapping breakpoint as produced by the csaw analysis was deemed broken and was noted in the resulting table. Arriba[20] (v1.1.0) was used to detect gene fusions using RNA-seq data, using the default settings which only outputs fusions with at least two supporting reads.
Grouping SVs at the patient level: To identify shared and unique SVs in patients with primary and relapse tumors, the per sample SV calls in BEDPE format were combined into one call-set as a pairs object using the R library rtracklayer (v1.46.0) and function import. Only high confidence SVs (called by 2 or more callers) were kept. A "GInteractions" object was made from the SVs using the R library InteractionSet (v1.14.0) and mode "strict". Then the "findOverlaps" method was used to find the interacting overlaps in the SV call set. The parameters used for "findOverlaps" were (maxgap = 5, ignore.strand=FALSE, select = "first", type = "start", use.region="both").

2.5 Copy number variant (CNV) annotation

CNVs were annotated for genes using the R package csaw (v1.20.0) against the GRCh37.87 annotation release using the parameters (dist=5000, promoter=c(3000, 1000), ignore.strand=TRUE). The VCF is available at: https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/common_all_20180423.vcf.gz. Segments were annotated to contain all genes overlapping the segments along with flanking parameters as described above. CNVs were categorized as follows: total copy number (TCN); loss of heterozygosity (LOH); lower or minor copy number (LCN). The classification of copy number alteration types is provided below in Supplementary Table 18. *CCNE1* copy number was previously assessed[53] by in situ hybridization (ISH) in 27 of the 126

primary tumors, with whole-genome (FACETS) derived *CCNE1* TCN showing a positive correlation with ISH derived *CCNE1* copy number (Spearman r = 0.55, *P* = 0.0025). A threshold of TCN ≥7 for gene amplification achieved a high specificity, with 96.3% (26/27) of tumors concordantly classified for *CCNE1* amplification status by FACETS and ISH.

Supplementary Table 18: Classification of copy number alteration types

Deletion (DEL)	TCN = 0
Loss (LOSS)	TCN = 1
Neutral	TCN = 2, LCN = 1
Gain (GAIN)	TCN > 2 and TCN ≤ 6
Amplification (AMP)	TCN ≥ 7
Copy-neutral loss-of-heterozygosity (cnLOH)	TCN = 2, LCN = 0

In order to visualize fluctuations in coverage associated with CNVs, bigWig tracks were generated on whole-genome sequencing data using deepTools (v3.0.0) with parameters (--binSize 10 --minMappingQuality 10 --normalizeUsing CPM --skipNonCoveredRegions --samFlagExclude 1024 --outFileFormat bigwig). Only chromosomes 1-22 and X were used for normalisation and track generation. Encode mappability blacklisted regions were excluded from the analysis.

2.6 Recurrent gene mutation detection

Small-scale mutations: The dNdScv[54] R package (v0.0.1.0) was used to detect putative driver genes enriched with small scale mutations (base substitutions and indels). High confidence SNVs and indels were fed into the "dndscv" function with parameters (refdb = "hg19", sm = "192r_3w", max_muts_per_gene_per_sample = Inf, max_coding_muts_per_sample = Inf, use_indel_sites = FALSE).

Structural variants: High confidence SVs in protein coding genes were selected for SV driver analysis. Breakpoints were analyzed using the tool GRIN[55] (v1.4). Results were further filtered to only contain expressed genes (those with a CPM greater than 0.5 in at least 10 samples) that were not in fragile sites, and *P* values were corrected again using GRIN's internal "robust.fdr" method.

Recurrent CNV analysis: Recurrent CNVs were identified using GISTIC2[56] (v2.0.23). LogR ratios in the segmentation data were first median centred per sample, then fed into the GISTIC2

tool with parameters (-savegene 1 -maxspace 1000 -ta 0.1 -td 0.1 -rx 0 -cap 3 -broad 0 -twoside 1 -res 0.05 -genegistic 0 -v 10).

Altered gene enrichment analysis: To compare the frequencies of gene mutations across survival groups, all types of high confidence variants were collated into one table and the following workflow was applied to generate the variants of interest per sample across the cohort:

1) A list of expressed protein coding genes was generated using RNA-seq data. Expressed genes were those with a CPM greater than 0.5 in at least 10 samples.
2) For gene level annotated CNVs with multiple overlapping segments, in order to select the most deleterious "loss-of-function" event, only the segment with the lowest logR value was kept.
3) For genes with high confidence SVs, if a gene had multiple breakpoints, only the breakpoint with the highest allele frequency was retained.
4) For high confidence small-scale variants (SNVs, indels and MNVs), only the most severe consequence variant as ranked by VEP was kept per gene. The order of severity is available on the Ensembl website: https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html
5) CNVs, SVs and small-scale variants were then combined and filtered to only retain the expressed protein coding genes as generated in 1).
6) For samples with genes affected by multiple variants, in order to select the most damaging variant, each gene was annotated per sample using the rank order shown below in Supplementary Table 19.
7) Fisher's test was performed to compare the proportion of mutated (high and moderate impact variants) and wild-type (low impact or no variant) samples for each survival group versus the other (i.e. LTS vs MTS, LTS vs STS, MTS vs STS).

Supplementary Table 19: Rank of gene alteration severity

Rank	Variant Event	Impact Level	Variant Type
1	Nonsense_Mutation	High	SNV, Indel, MNV
2	Frame_Shift_Del	High	SNV, Indel, MNV
3	Frame_Shift_Ins	High	SNV, Indel, MNV
4	Splice_Site	High	SNV, Indel, MNV
5	CNV_Del = 0	High	CNV

6	SV_gene_body	High	SV
7	Translation_Start_Site	High	SNV, Indel, MNV
8	Nonstop_Mutation	High	SNV, Indel, MNV
9	CNV_Amp = ≥7	High	CNV
10	Missense_Mutation	Moderate	SNV, Indel, MNV
11	In_Frame_Ins	Moderate	SNV, Indel, MNV
12	In_Frame_Del	Moderate	SNV, Indel, MNV
13	Splice_Region	Low	SNV, Indel, MNV
14	Silent	Low	SNV, Indel, MNV
15	None	None	Any

Filtering of altered genes of interest: Genes were filtered as follows: 1) to include only expressed genes, that is genes with > 0.5 CPM in at least 10 samples, 2) to remove 49 genes heavily affected by artefacts listed here[54], 3) to remove FrequentLy mutAted GeneS (FLAGS[57]) listed in section "Whole Genome Sequencing (Somatic SNV/Indel calling)", and 4) to remove genes affected by fragile sites as previously identified in the PCAWG studies[58,59]. Cancer-related genes were those included in the COSMIC Cancer Gene Census version "Thu Mar 26 00_38_30 2020" (https://cancer.sanger.ac.uk/census).

2.7 Mutual exclusivity analysis

Pairwise Fisher's exact tests for mutual exclusivity or co-occurrence of alterations in genes of interest was performed using the "somaticInteractions" function of maftools[60] (v2.2.10), only including genes with alterations in a minimum of three primary tumors. *P* values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure.

2.8 Clonality analysis

DNA repair pathway alterations were manually reviewed in IGV to get the most accurate estimate of their depth and variant read count. The alteration positions were then intersected with FACETS copy number data to determine the total copy number at each variant locus. The depth, variant read counts and FACETS total copy number for each alteration, along with FACETS purity estimates for each sample, were fed into the function "absolute.cancer.cell.fraction", which calculates the purity and copy number corrected cancer cell fraction (CCF) and classifies the clonal or subclonal status of each variant. The function is

freely available at the GitHub repository [https://github.com/ucl-respiratory/preinvasive], and the methods to calculate the CCF and classify clonality are described in the associated publication[12].

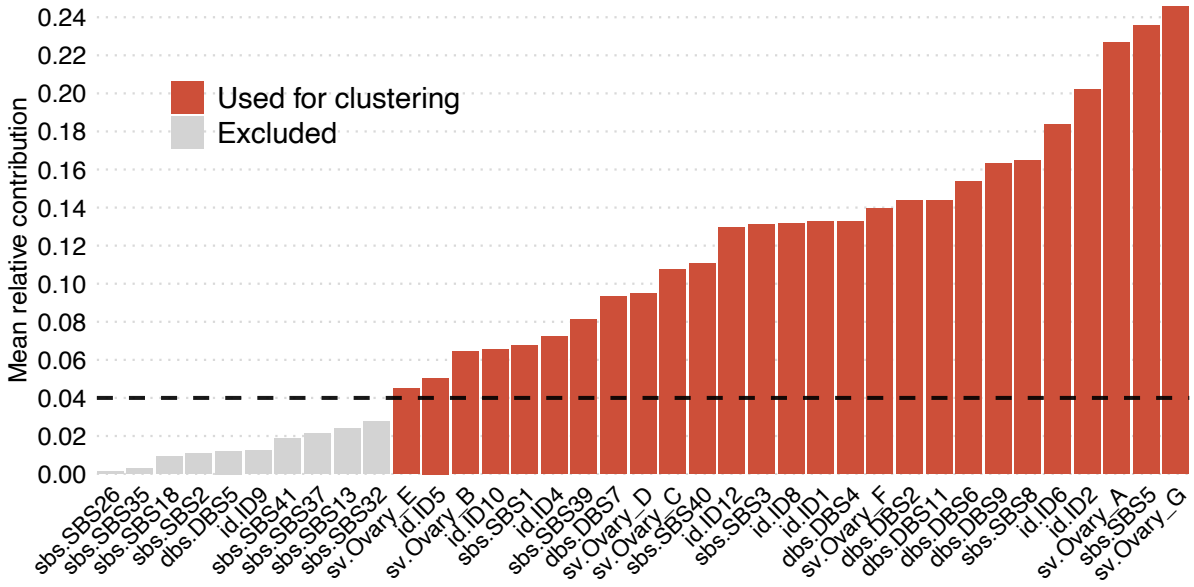
To estimate clonality of copy number segments containing *CCNE1*, the FACETS cellular fractions (CF) of *CCNE1* segments were converted to cancer cell fractions (CCF) by dividing the CF by the sample purity (Supplementary Table 17). A CF value of 1 in FACETS indicates normal ploidy, so segments with a CF of 1 were designated a CCF of 0. CCF values were upper bound to 1 to account for noise. All *CCNE1* amplifications were in copy number segments with CCFs in the range of 0.999 to 1 and were considered clonal.

2.9 Mutational signature fitting

Signature fitting was performed using the R package signature.tools.lib and its function "SignatureFit_withBootstrap" [method = "KLD", nboot = 100, randomSeed = 42, threshold_percent = 2, threshold_p.value = 0.05]. Fitting was carried out in two steps. An initial signature fitting was performed using all single base substitution (SBS), double base substitution (DBS) and indel (ID) signatures in COSMIC (https://cancer.sanger.ac.uk/signatures/), and SV signatures in SIGNAL (https://signal.mutationalsignatures.com/)[27]. This generated absolute signature exposures or contributions per sample, per signature. Then a second fit was performed using just the ovary identified signatures in COSMIC (for SBS, DBS and ID) and SIGNAL (for SV), along with any other signatures that were present at high levels in the cohort. This was determined by visualising the mean of the signature exposures across the cohort and selecting an acceptable threshold. The resulting absolute contributions could then be converted into relative contributions by dividing each of the absolute contributions by the sum total per sample.

2.10 Mutational signature clustering

The relative contributions of the COSMIC (SBS, DBS, ID) and SIGNAL (SV) signatures fit by signature.tools.lib were used to generate the mutational signature clusters. The sample by signature proportion matrix was filtered to only contain signatures with a mean of at least 4% across the cohort. This was determined by visualising the mean contribution of each signature across the cohort and selecting an appropriate cut-off (Supplementary Fig. 20).



Supplementary Fig. 20 | Selection of mutational signatures used for clustering. Bars indicate the mean relative signature contribution (y axis) across the cohort of primary tumor samples (n = 126), of the 37 COSMIC and SIGNAL signatures fit by signature.tools.lib. Dotted line indicates the cut-off and red bars indicate signatures that were included in the clustering. Signatures are ordered by mean relative contribution in ascending order.

The proportions were then scaled and centred across the cohort using the "scale" function in R. The standardized values were fed into the R package ConsensusClusterPlus[61] (v1.50.0) with parameters (maxK = 10, reps=1000, pItem=0.9, pFeature=0.9, clusterAlg="pam", distance="pearson", innerLinkage="ward.D2", finalLinkage="ward.D2", seed=12345678) to generate sample clusters. The heatmap and cumulative distribution function (CDF) plots of the consensus matrices, and the Delta Area Plot generated by ConsensusClusterPlus was visualized to select an appropriate cut-off for the number of clusters. The chosen cut-off was also based upon visual inspection of the final sample clusters with their known features whilst retaining an appropriate cluster size for statistical analyses. The rows (signatures) were clustered using R functions dist (method = "euclidean") and hclust (method = "ward.D2"). Both sample and signature clusters were reordered for optimal leaf ordering using the function "reorder.hclust" from the R package "seriation" (v1.3.0) with parameter (method = "OLO").

2.11 RNA sequencing batch correction

The batch correction was performed in order to remove batch effects while retaining group differences using limma's removeBatchEffect function with the parameters (exp_data, batch = LibraryType, design = model.matrix(~SurvivalGroup)) where "exp_data" is the $\log_2$ TMM normalized data. The design of the study is shown in Supplementary Table 20 below.

Supplementary Table 20: RNA-seq samples grouped by library type and survival group

SurvivalGroup	LibraryType stranded	LibraryType unstranded
LTS	56	7
STS and MTS	0	68

Given the unbalanced nature of the RNA sequencing data, we acknowledge that the batch corrected expression data may have some limitations and points of caution[62]. When using batch corrected expression data produced by limma for statistical inference, true biological differences may be exaggerated. On the other hand, these same differences might be dampened by tools such as DESeq2 that incorporate the batch into the model itself. Where possible when performing comparative or differential expression analysis between the survival groups for individual genes, both the boxplots of the normalized and batch corrected expression data and the DESeq2 generated adjusted *P* values and fold-changes are shown. The DeepCC expression analysis is largely unaffected by such biases, as stated by the authors[63].

2.12 Expression analyses

Molecular subtype classification: DeepCC[63] (v0.1.1) was used for the molecular subtype classification of tumor samples. The Tothill[42] dataset of ovarian tumor samples (GSE9891) was used for training. Of the 285 samples in the dataset, 215 C1, C2, C4 and C5 HGSC samples were used to generate the classifier. The DeepCC function "getFunctionalSpectra" with default parameters was used on the Tothill expression matrix to get the functional spectra. The function "train_DeepCC_model" was used to train the model. Since the training gave slightly different results in each iteration, 10 iterations were performed to get 10 different models. Each of the models were then run against the functional spectra of our batch corrected RNA-seq data to obtain subtype predictions. The subtype predicted the greatest number of times, after classifying using each of the 10 models, was selected as the final subtype.

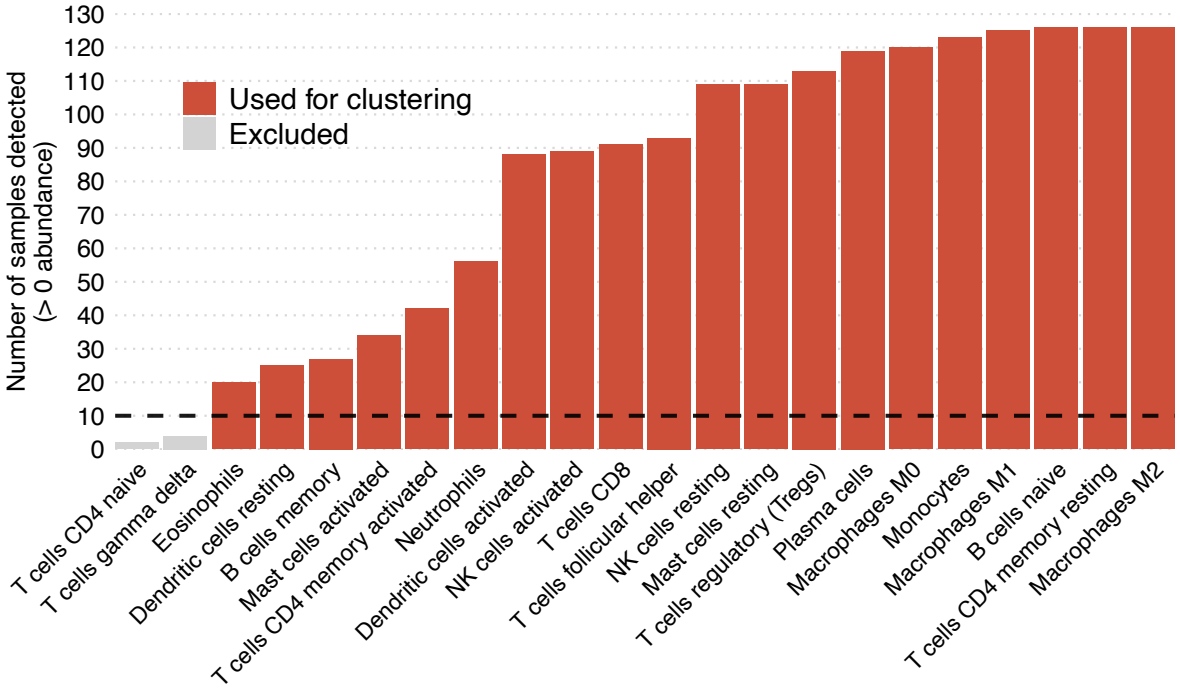
Differential expression analysis: DESeq2[64] (v1.26.0) was used to identify differentially expressed protein coding genes between sample groups of interest, with batch effects accounted in the model. Raw counts were filtered to remove lowly expressed genes before running the analysis.

Gene Set Enrichment Analysis (GSEA): An R package for fast, pre-ranked gene set enrichment analysis (FGSEA v1.15.1) was used to calculate gene set enrichment across comparison groups. *P* values obtained from DESeq2 were transformed to signed *P* values by converting the gene level *P* values to a negative $\log_{10}$ value and applying the sign of the fold change to these values. These values were then sorted and fed into FGSEA via its function fgseaMultilevel (minSize=15, maxSize = 500, gseaParam = 0, eps = 0) to generate enrichment scores and FDR adjusted *P* values across the Hallmark gene sets in the MSigDB database[49] (v7.4).

Immune cell deconvolution: CIBERSORTx[28] was used to estimate immune cells present in the tumor microenvironment using the web-based pipeline (https://cibersortx.stanford.edu/). The analysis was run with batch correction enabled and B-mode, quantile normalisation disabled, absolute mode and 500 permutations. The LM22 signature matrix was used for immune cell deconvolution.

CIBERSORTx survival analysis: CIBERSORTx absolute values were centred and scaled across the cohort using the R function "scale". A multi-variate Cox analysis for overall survival using all detected cell types was performed using the R package survival (v3.2-7) and its function "coxph" with parameter (robust = TRUE).

Immune cell clustering: CIBERSORTx absolute cell abundances were used to generate immune clusters. Only cell types with a value greater than 0 in at least 10 samples were used (Supplementary Fig. 21). The sample by cell type abundance matrix was then used to generate immune clusters, as per the method described for the mutational signature cluster analysis.



Supplementary Fig. 21 | Selection of cell types used for clustering. Bars indicate the number of primary tumor samples (y axis) for which the CIBERSORTx cell type abundance was greater than 0. Dotted line indicates the cut-off and red bars indicate cell types that were included in the clustering. Cell types are ordered by number of tumor samples with abundance >0 in ascending order.

2.13 Methylation analyses

Differential methylation (DM) analysis: A shared methylation array probe set (364,406 probes) was generated for all 126 primary tumor samples as described in "Methylation data processing and quality control" in the Methods. The DM analysis between survival groups was performed following the workflow described here: https://www.bioconductor.org/packages/release/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html. Multi-dimensional scaling (MDS) plots were generated to visualize potential sources of variation between samples. Methylation array specific variation was observed between samples run on the EPIC versus the 450k arrays, and therefore was added to the statistical model of limma for DM analysis. To identify methylation sites that negatively correlate with expression, a Pearson correlation analysis was performed between batch corrected RNA-seq data and methylation beta values for 126 primary tumor samples. Correlation analysis was performed on each probe and its associated gene using the

base R “cor.test” function. DM results for each probe in each survival group comparison was then annotated with expression specific correlation analysis data as described above.

Only CpG probes with an unadjusted *P* value <0.05 were output to results tables (Supplementary Data 6). A column was added to each DM analysis results table to indicate the genes that were deemed to be turned off in the long-term survivors (EXP_TURNED_OFF_IN_LTS) versus the short- or moderate-term survivors, and the genes that were deemed to be turned off in the moderate-term survivors (EXP_TURNED_OFF_IN_MTS) in contrast to the short-term survivors. The filters used were as follows: 1) adjusted *P* value < 0.1 to select for significant probes, 2) logFC > 0 to select for probes with higher methylation values in either the long- or moderate-term survivors in contrast to the relevant comparison group, and 3) Pearson correlation with expression data < -0.2 to select for probes that negatively correlated with gene expression in the entire cohort of 126 primary tumors.

Methylation clustering: Unsupervised consensus clustering of methylation data was performed across all primary tumors (*n* = 126) using the top 1% most variable CpG probes[65] across the cohort (number of probes = 3,645). Batch corrected methylation beta values were transformed to M-values[66], and clustering was performed with M-values using the clustering method described above for the mutational signature cluster analysis. The ConsensusClusterPlus parameters used for the methylation data clustering were: maxK = 10, reps=1000, pItem=0.9, pFeature=1, clusterAlg="pam", distance="euclidean", innerLinkage="ward.D2", finalLinkage="ward.D2", seed=12345678.

Gene promoter methylation classification: BRCA1 and *RAD51C* promoter methylation status was determined by comparing *BRCA1* and *RAD51C* mRNA expression to the methylation array beta values of probes located in the 5’ CpG island of each respective gene across primary and relapse tumor samples (*n* = 131), following previously established approaches[10,14]. Briefly, methylation probes demonstrating a negative correlation with mRNA expression (Spearman’s r <-0.30, P < 0.01) were selected (*BRCA1* probes: cg04110421, cg04658354, cg08386886, cg09441966, cg09831010, cg10893007, cg15419295, cg16630982, cg16963062, cg19088651, cg20187250, cg21253966, cg24806953; *RAD51C* probes: cg02118635, cg05214530, cg24099023), and samples with a beta value of >0.2, in at least 50% of selected probes, were classified as methylated.

2.14 Multivariable analysis

Univariable associations with progression-free and overall survival was determined using Cox proportional-hazards regression models for all features of interest (Fig. 6b; Supplementary Table 13). Continuous features (T cells CD4 memory activated, *PCNA* expression, plasma cells, neoantigen count, NK cells) were first scaled and centred using the R function “scale” before analysis using the “coxph” function. Features significantly associated with overall survival ($P < 0.05$) were selected for the multivariable analysis to ascertain the most significant independent features of prognostication (Supplementary Table 14). A pair-wise correlation analysis between all prognostic factors was used to investigate their independence using the R library polycor (v0.8-1) and its function "hetcor", which accounts for different combinations of continuous and categorical variables.

Supplementary References

1 Kurman, R. J., Carcangiu, M. L., Herrington, C. S., Young, R. H. & World Health Organization. *WHO Classification of Tumours of Female Reproductive Organs*. (IARC Press, 2014).

2 Silverberg, S. G. Histopathologic Grading of Ovarian Carcinoma: A Review and Proposal. *Int. J. Gynecol. Pathol.* **19** (2000).

3 Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 1-12, doi:10.1038/s41467-020-19406-4 (2020).

4 Moghadasi, S. *et al.* The BRCA1 c. 5096G>A p.Arg1699Gln (R1699Q) intermediate risk variant: breast and ovarian cancer risk estimation and recommendations for clinical management from the ENIGMA consortium. *J. Med. Genet.* **55**, 15-20, doi:10.1136/jmedgenet-2017-104560 (2018).

5 Joshi, P. M., Sutor, S. L., Huntoon, C. J. & Karnitz, L. M. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *J. Biol. Chem.* **289**, 9247-9253, doi:10.1074/jbc.M114.551143 (2014).

6 Bajrami, I. *et al.* Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Research* **74**, 287-297, doi:10.1158/0008-5472.CAN-13-2541 (2014).

7 Popova, T. *et al.* Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. *Cancer Res.* **76**, 1882-1891, doi:10.1158/0008-5472.CAN-15-2128 (2016).

8 Wu, Y. M. *et al.* Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. *Cell* **173**, 1770-1782.e1714, doi:10.1016/j.cell.2018.04.034 (2018).

9 Maxwell, K. N. *et al.* BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nat. Commun.* **8**, 319-319, doi:10.1038/s41467-017-00388-9 (2017).

10 The Cancer Genome Atlas Research Network. Integrated genomic analysis of ovarian cancer. *Nature* **474**, 609-615 (2011).

11 Etemadmoghadam, D. *et al.* Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc. Nat. Acad. Sci. USA* **110**, 19489-19494, doi:10.1073/pnas.1314302110 (2013).

12 Teixeira, V. H. *et al.* Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nat. Med.* **25**, 517-525, doi:10.1038/s41591-018-0323-0 (2019).

13 Ahmed, A. A. *et al.* Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *J. Pathol.* **221**, 49-56, doi:10.1002/path.2696 (2010).

14 Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489-494, doi:10.1038/nature14410 (2015).

15 Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, doi:10.1038/s41586-020-1969-6 (2020).

16 Anaya, J. OncoLnc: Linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Computer Science* **2016**, doi:10.7717/peerj-cs.67 (2016).

17 Bronger, H. *et al.* CXCL9 and CXCL10 predict survival and are regulated by cyclooxygenase inhibition in advanced serous ovarian cancer. *Br. J. Cancer* **115**, 553-563, doi:10.1038/bjc.2016.172 (2016).

18 Millstein, J. *et al.* Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann. Oncol.* **31**, 1240-1250, doi:10.1016/j.annonc.2020.05.019 (2020).

19 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

20 Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448-460, doi:10.1101/GR.257246.119 (2021).

21 Salzman, J. *et al.* ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. *PLoS Biol.* **9**, e1001156-e1001156, doi:10.1371/journal.pbio.1001156 (2011).

22 Kannan, K. *et al.* CDKN2D-WDFY2 is a cancer-specific fusion gene recurrent in high-grade serous ovarian carcinoma. *PLoS Genet.* **10**, e1004216-e1004216, doi:10.1371/journal.pgen.1004216 (2014).

23 Newtson, A., Reyes, H., Devor, E. J., Goodheart, M. J. & Bosquet, J. G. Identification of Novel Fusion Transcripts in High Grade Serous Ovarian Cancer. *Int. J. Mol. Sci.* **22**, doi:10.3390/ijms22094791 (2021).

24 Kim, P. & Zhou, X. FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Res.* **47**, D994-D1004, doi:10.1093/nar/gky1067 (2019).

25 Garsed, D. W. *et al.* Homologous recombination DNA repair pathway disruption and retinoblastoma protein loss are associated with exceptional survival in high-grade serous ovarian cancer. *Clin. Cancer Res.* **24**, 569-580, doi:10.1158/1078-0432.CCR-17-1621 (2018).

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

27 Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat. Cancer* **1**, 249-263, doi:10.1038/s43018-020-0027-5 (2020).

28 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773-782, doi:10.1038/s41587-019-0114-2 (2019).

29 Nersesian, S. *et al.* NK cell infiltration is associated with improved overall survival in solid cancers: A systematic review and meta-analysis. *Transl. Oncol.* **14**, 100930-100930, doi:10.1016/j.tranon.2020.100930 (2021).

30 Henriksen, J. R. *et al.* Favorable prognostic impact of Natural Killer cells and T cells in high-grade serous ovarian carcinoma. *Acta Oncol.* **59**, 652-659, doi:10.1080/0284186X.2019.1711173 (2020).

31 Matsuda, T., Kawanishi, M., Yagi, T., Matsui, S. & Takebe, H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res.* **26**, 1769-1774 (1998).

32 Bobo, J. K. & Husten, C. Sociocultural influences on smoking and drinking. *Alcohol Res. Health* **24**, 225-232 (2000).

33 Wang, Y. *et al.* The effects and mechanisms of SLC34A2 in tumorigenesis and progression of human non-small cell lung cancer. *J. Biomed. Sci.* **22**, 1-12, doi:10.1186/s12929-015-0158-7 (2015).

34 Ye, W. *et al.* Overexpression of SLC34A2 is an independent prognostic indicator in bladder cancer and its depletion suppresses tumor growth via decreasing c-Myc expression and transcriptional activity. *Cell Death Dis.* **8**, 1-11, doi:10.1038/cddis.2017.13 (2017).

35 Soares, I. C. *et al.* In silico analysis and immunohistochemical characterization of NaPi2b protein expression in ovarian carcinoma with monoclonal antibody Mx35.

Appl. Immunohistochem. Mol. Morphol. **20**, 165-172, doi:10.1097/PAI.0b013e318228e232 (2012).

36 Moore, K. N. *et al.* Phase 1b study of anti-NaPi2b antibody-drug conjugate lifastuzumab vedotin (DNIB0600A) in patients with platinum-sensitive recurrent ovarian cancer. *Gynecol. Oncol.* **158**, 631-639, doi:10.1016/j.ygyno.2020.05.039 (2020).

37 Richardson, D. L. *et al.* Uplift (ENGOT-ov67): A pivotal cohort to evaluate XMT-1536 (upifitamab rilsodotin), a NaPi2b-directed antibody drug conjugate for platinum-resistant ovarian cancer. *J. Clin. Oncol.* **39**, TPS5607-TPS5607, doi:10.1200/JCO.2021.39.15_suppl.TPS5607 (2021).

38 Ge, G. *et al.* Enhanced SLC34A2 in breast cancer stem cell-like cells induces chemotherapeutic resistance to doxorubicin via SLC34A2-Bmi1-ABCC5 signaling. *Tumour Biol.* **37**, 5049-5062, doi:10.1007/s13277-015-4226-0 (2016).

39 Zhang, Z. *et al.* EPHA7 mutation as a predictive biomarker for immune checkpoint inhibitors in multiple cancers. *BMC Med.* **19**, 26-26, doi:10.1186/s12916-020-01899-x (2021).

40 Shukla, V. *et al.* BRCA1 affects global DNA methylation through regulation of DNMT1. *Cell Res.* **20**, 1201-1215, doi:10.1038/cr.2010.128 (2010).

41 Holm, K. *et al.* An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Res.* **18**, 27-27, doi:10.1186/s13058-016-0685-5 (2016).

42 Tothill, R. W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198-5208, doi:10.1158/1078-0432.CCR-08-0196 (2008).

43 Talhouk, A. *et al.* Development and Validation of the Gene Expression Predictor of High-grade Serous Ovarian Carcinoma Molecular SubTYPE (PrOTYPE). *Clin. Cancer Res.* **26**, 5411-5423, doi:10.1158/1078-0432.CCR-20-0103 (2020).

44 McCartney, D. L. *et al.* Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214-220, doi:10.1016/j.ebiom.2018.10.051 (2018).

45 Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622, doi:10.1126/science.aag0299 (2016).

46 Wilhelm-Benartzi, C. S. *et al.* Association of secondhand smoke exposures with DNA methylation in bladder carcinomas. *Cancer Causes Control* **22**, 1205-1213, doi:10.1007/s10552-011-9788-6 (2011).

47 Tsai, P.-C. *et al.* Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clin. Epigenetics* **10**, 126-126, doi:10.1186/s13148-018-0558-0 (2018).

48 Etemadmoghadam, D. *et al.* Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. *Clin. Cancer Res.* **15**, 1417-1427, doi:10.1158/1078-0432.CCR-08-1564 (2009).

49 Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

50 Shen, R. & Seshan, V. E. FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, 1-9, doi:10.1093/nar/gkw520 (2016).

51 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1-4, doi:10.1093/gigascience/giab008 (2021).

52 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1-14, doi:10.1186/s13059-016-0974-4 (2016).

53 Aziz, D. *et al.* 19q12 amplified and non-amplified subsets of high grade serous ovarian cancer with overexpression of cyclin E1 differ in their molecular drivers and clinical outcomes. *Gynecol. Oncol.* **151**, 327-336, doi:10.1016/j.ygyno.2018.08.039 (2018).

54 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823-1823, doi:10.1016/j.cell.2018.06.001 (2018).

55 Pounds, S. *et al.* A genomic random interval model for statistical analysis of genomic lesion data. *Bioinformatics* **29**, 2088-2095, doi:10.1093/bioinformatics/btt372 (2013).

56 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41-R41, doi:10.1186/gb-2011-12-4-r41 (2011).

57 Shyr, C. *et al.* FLAGS, frequently mutated genes in public exomes. *BMC Med. Genomics* **7**, 64-64, doi:10.1186/s12920-014-0064-y (2014).

58 Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, doi:10.1038/s41586-020-1965-x (2020).

59 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121, doi:10.1038/s41586-019-1913-9 (2020).

60 Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747-1756, doi:10.1101/gr.239244.118 (2018).

61 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573, doi:10.1093/bioinformatics/btq170 (2010).

62 Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29-39, doi:10.1093/biostatistics/kxv027 (2016).

63 Gao, F. *et al.* DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, doi:10.1038/s41389-019-0157-8 (2019).

64 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550-550, doi:10.1186/s13059-014-0550-8 (2014).

65 The Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e928, doi:10.1016/j.cell.2017.10.014 (2017).

66 Xie, C. *et al.* Differential methylation values in differential methylation analysis. *Bioinformatics* **35**, 1094-1097, doi:10.1093/bioinformatics/bty778 (2019).