The genomic landscape of early-stage ovarian high grade serous carcinoma

Zhao Cheng[1], Hasan Mirza[1]*, Darren P. Ennis[1]*, Philip Smith[2], Lena Morrill Gavarró[2], Chishimba Sokota[3], Gaia Giannone[1,4], Theodora Goranova[2], Thomas Bradley[2], Anna Piskorz[2]; Michelle Lockley[5] for the BriTROC-1 Investigators^; Baljeet Kaur[3], Naveena Singh[6], Laura A. Tookman[1], Jonathan Krell[1], Jacqueline McDermott[7]; Geoffrey Macintyre[2], Florian Markowetz[2], James D. Brenton[2] and Iain A. McNeish[1+]

1. Ovarian Cancer Action Research Centre, Department of Surgery and Cancer, Imperial College London, London, UK
2. Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
3. Department of Cellular Pathology, Imperial College Healthcare NHS Trust, London, UK
4. Department of Oncology, University of Turin, Turin, Italy
5. Centre for Cancer Cell and Molecular Biology, Barts Cancer Institute, Queen Mary University of London, London, UK
6. Department of Pathology, Barts Healthcare NHS Trust, London, UK
7. Department of Pathology, University College London Hospital NHS Trust, London, UK

* equal contribution
^ The BriTROC-1 Investigators are listed in the Supplementary Acknowledgements

Running title:
Genomics of early stage ovarian high grade serous carcinoma

[+]: To whom correspondence should be addressed:
Professor Iain McNeish
Imperial College London
IRDB Building
Hammersmith Hospital
W12 0NN
Tel: +44 20 7594 2792
i.mcneish@imperial.ac.uk

Conflict of interest statement:

The authors have declared no conflicts of interest relating to the work presented here.

Statement of translational relevance

To determine whether early-stage ovarian high grade serous carcinoma (HGSC) represents a distinct genomic entity, we collected samples from 43 patients with stage IA-IIA HGSC to identify potential differences in short genomic variants and copy number aberrations, and compared them to a cohort of 52 late-stage (stage IIIC-IV) cases. We found no significant differences in somatic mutations or focal copy number alterations between early-stage and late-stage cohorts. There was, however, a significant difference in both ploidy and copy number signature exposure between early and late-stage samples, with higher ploidy and signature 4 exposure in late-stage cases. Unsupervised hierarchical clustering revealed three clusters, which were prognostic. Together, our data suggest that early and late-stage HGSC share fundamental genomic features, but that late-stage disease appears distinct from early-stage, with evidence of whole genome duplication that may provide evolutionary benefit.

Abstract

Purpose: Ovarian high grade serous carcinoma (HGSC) is usually diagnosed at late stage. We investigated whether late-stage HGSC has unique genomic characteristics consistent with acquisition of evolutionary advantage compared to early-stage tumours.

Experimental Design: We performed targeted next generation sequencing and shallow whole genome sequencing (sWGS) on pre-treatment samples from 43 patients with FIGO stage I–IIA HGSC to investigate somatic mutations and copy number alterations (SCNA). We compared results to pre-treatment samples from 52 stage IIIC/IV HGSC patients from the BriTROC-1 study.

Results: Age of diagnosis did not differ between early-stage and late-stage patients (median 61.3 years vs 62.3 years respectively). *TP53* mutations were near-universal in both cohorts (89% early-stage, 100% late-stage) and there were no significant differences in the rates of other somatic mutations, including *BRCA1* and *BRCA2*. We also did not observe cohort-specific focal SCNA that could explain biological behaviour. However, ploidy was higher in late-stage (median 3.0) than early-stage (median 1.9) samples. Copy number (CN) signature exposures were significantly different between cohorts, with greater relative signature 3 exposure in early-stage and greater signature 4 in late-stage. Unsupervised clustering based on CN signatures identified three clusters that were prognostic.

Conclusions: Early stage and late stage HGSC have highly similar patterns of mutation and focal SCNA. However, copy number signature analysis showed that late-stage disease has distinct signature exposures consistent with whole genome duplication. Further analyses will be required to ascertain whether these differences reflect genuine biological differences between early and late-stage or simply time-related markers of evolutionary fitness.

Introduction

High grade serous carcinoma (HGSC) accounts for approximately 70% of all ovarian cancer (OC) cases and approximately 80% of OC deaths. Most patients with HGSC present with advanced (FIGO stage III and IV) disease, where treatment is rarely curative. Despite the addition of anti-angiogenic agents (1,2) and PARP inhibitor therapy (3,4), the majority of patients with advanced disease relapse within 24 months of completion of first-line chemotherapy. By contrast, the small proportion (10 - 15%) patients who present with early disease (stage I and II) have much better prognosis and are frequently cured with surgery and platinum-based chemotherapy alone (5).

HGSC is marked by near-universal *TP53* mutation (6,7) and arises from the fimbriae of the distal fallopian tube, evolving from p53 signatures (cytologically normal cells with mutant *TP53*), via serous intra-epithelial carcinomas (STIC) to invasive carcinomas (8) that metastasise to the ovary and throughout the peritoneal cavity. HGSC is marked by widespread copy number change and is the archetypal C class, copy number-driven malignancy (9). Although the genome of HGSC is highly complex, we recently described copy number signatures, recurrent patterns of genome-wide copy number change that were prognostic and were significantly associated with specific driver mutational processes (10).

The large studies that defined the genomic landscape of HGSC, including those from The Cancer Genome Atlas consortium (TCGA) (7), the Australian Ovarian Cancer Study (11) and the International Cancer Genome Consortium (ICGC) (12), all analysed samples from patients almost exclusively with stage III or IV disease and there is little information about the genomics of early-stage HGSC. Previous analyses of genomic alterations in matched p53 signatures, STIC, ovarian and metastatic lesions in patients with late-stage HGSC have shown clearly that there are common SNV/indel in all samples from each patient (13), and that SCNA are evident very early in disease development and are shared across samples, with an estimated seven years required to progress from first mutation to clinical presentation with advanced disease (14).

However, it is unclear whether patients with early-stage HGSC have a distinct subtype of disease that fails to metastasise or whether these cases are genomically similar to late-stage disease but are identified essentially by chance before metastasising. To address this, we have undertaken genomic analysis, including shallow whole genome sequencing and deep sequencing of a target gene panel, of a cohort of early-stage HGSC patients identified at three large UK centres, with comparison to late-stage samples from the BriTROC-1 study.

Patients and Methods

Study conduct, survival analyses and patient samples

Details of the BriTROC-1 study have been reported previously (10,15) – the study was conducted in accordance with the principles of the Declaration of Helsinki. Ethics/IRB approval was given by Cambridge Central Research Ethics Committee (Reference 12/EE/0349) and all patients gave written informed consent to participate. For all other cases, formalin-fixed paraffin-embedded samples were obtained from the pathology archives of participating hospitals by specialist gynaecological pathologists (BK, NS, JMcD) and utilised under the auspices and ethical approval of the Imperial College Healthcare Tissue Bank (HTA licence 12275, Research Ethics Committee number 17/WA/0161, Project ID R18060). Inclusion criteria were stage IA, IB, IC or IIA ovarian or fallopian tube carcinomas of high-grade serous histology diagnosed in the previous 15 years, and patients identified through routine clinical practice. We utilised the FIGO staging criteria in use at the time of diagnosis. Exclusion criteria included non-high grade serous pathology and samples identified at risk-reducing surgery. All samples were obtained prior to any chemotherapy treatment – the BriTROC-1 samples were those obtained at the time of diagnosis rather than relapse. At Imperial, a database of 680 patients initially revealed 45 patients with stage I – IIA high grade serous carcinoma. However, only 20 patients fulfilled all inclusion/exclusion criteria. Overall survival was calculated from the date of diagnosis to the date of death or the last known clinical assessment. All cases, including the early-stage cases from BriTROC-1, underwent pathological review (CS, JMcD).

Sequencing

Details of the sequencing of BriTROC-1 samples are given elsewhere (10). For new samples, DNA was extracted from 10 × 10 µm sections using QIAmp DNA FFPE Tissue Kit (Qiagen, UK) according to the manufacturer's protocol. 50-200ng was sheared with a Covaris LE220 focused-ultrasonicator (Covaris, Woburn, MA) to produce 100-200bp fragments. Libraries were generated using SureSelect XT standard protocol (Agilent Technologies, Santa Clara, CA) for low-input and FFPE samples. Analysis of *PTEN, KRAS, RB1, BRCA2, RAD51B, FANCM, PALB2, RAD51D, TP53, RAD51C, BRIP1, CDK12, NF1, BRCA1, BARD1, PIK3CA* was performed using a custom Ampliseq panel on a HiSeq4000 system (Illumina, Cambridge, UK), using paired-end 125 bp protocols. The mean coverage was >7000×. Nine samples were used as a panel of normal controls, 5 of them adjacent normal tissue and 4 samples whole blood. Shallow whole genome sequencing (sWGS) was performed on a HiSeq4000 system (Illumina Cambridge, UK), using paired-end 150 bp protocols, with 250-300 ng input DNA according to the manufacturer's instructions. The minimum number of reads per sample was set at 5-10 million (mean coverage of 0.1×). Using our previous calculations (https://gmacintyre.shinyapps.io/sWGS_power/), 10 million

reads with a bin size of 30kb had 80% power (a = 0.01) to detect CN change +/- 2 at 30% purity assuming ploidy of 2.

Mutation calling

FASTQ files were trimmed for adapters and aligned to reference human genome hg19 using Burrows-Wheeler Alignment (BWA-MEM) (16) and pre-processed using samtools and Picard to generate sorted BAM files (17). Somatic mutations were called using Mutect2 (GATK4.1.4.1) (18), Varscan2 (version 2.4.2) (19), Strelka2 (version 1.0.14) (20) and HaplotypeCaller (21) pipelines for single nucleotide variations (SNVs) and small insertions and deletions (indels) on tumour-only BAM files (new early stage samples) and tumour-normal pairs for the BriTROC-1 samples, where available, using default parameters. Where multiple samples existed for individual BriTROC-1 patients, sample data in sorted BAM files were merged prior to mutation calling but individual sample identity was retained. Mutations were annotated using Variant Effect Predictor (VEP) (version 1.5.3) (22). Somatic mutations were filtered by clinical significance (ClinVar, October 2020) with “benign” and “likely_benign” variants discarded. Variants were further checked for pathogenicity in the COSMIC database (GRCh37, February 2021). Germline mutations for BriTROC-1 samples were detected using similar criteria for pathogenicity using Strelka and HaplotypeCaller.

Absolute copy number and copy number signature calling

sWGS reads were aligned to reference human genome hg19. Relative copy numbers were obtained for predefined 30kb bins using a modified version of the QDNASeq package (23). We obtained absolute copy numbers using the sWGS-absoluteCN (swgs) pipeline - full details are given in Supplementary Information. Focal amplifications and deletions were defined according to the COSMIC definitions (see https://cancer.sanger.ac.uk/cosmic/help/cnv/overview): amplification was defined as total copy number ≥5 if average ploidy ≤2.7, or ≥9 if average ploidy >2.7. Loss was defined as total copy number 0 if average ploidy ≤2.7 or (ploidy minus 2.7) if average ploidy >2.7. Gain in Figure 3C was defined as total copy number >2.5 but <5.0. Copy number signatures were calculated using the R scripts as previously published (10).

Copy number signature comparison

To model the presence or absence of signatures, a fixed effects Bernoulli model was used with an intercept and a coefficient for the change between early and late-stage samples. The presence of a signature j in sample i is modelled by a Bernoulli with probability θ_{ij}, where $\theta = x^{T}\beta$. x has two

rows – for the intercept and the difference between the groups – and as many rows as samples. β has two rows and as many columns as the number of signatures (d = 7). The change in the differential abundance of non-zero exposures has been modelled similarly. We used a multivariate normal model based on the isometric log ratio (ILR)-transformed exposures (also called a logistic-normal model in the literature). The ILR transformation maps a d-dimensional compositional vector (the exposures) to a d−1 dimensional vector of real values. To account for absent signatures, the transformation (and subsequently the model) only used the subset of signatures that are present in each sample. Covariates are the same as those in Bernoulli model, but this time $\theta = x^T\beta$ represents the ILR-transformed probabilities. Therefore, it has 6 columns instead of 7, and each row can be transformed back to a 7-dimensional vector of probabilities with the inverse ILR transformation. β continues to have two rows, but only 6 columns, which indicate changes in log-ratios of signature exposures. The R package TMB (24) was used for inference. The model was written in C++ and a full description of the analysis is given in Supplementary Methods.

Unsupervised clustering of patients using signature exposures

Hierarchical clustering of the copy number signature exposure vectors of all samples (early-stage and late-stage) used in the survival analysis was performed using the NbClust (25) package in R. The NbClust package contains 30 indices to determine the relevant number of clusters; the number of 3 clusters ranked the top clustering scheme from different results obtained by varying all combinations of number of clusters, distance measures and clustering methods. A Cox proportional hazards model was fitted using the cluster labels as covariates, using the R packages survival (26) and survminer (https://rpkgs.datanovia.com/survminer/index.html). For survival analyses based on cluster, patients with >1 sample were allocated into the cluster of the sample with highest purity.

Data accessibility

All sequencing data are available via the European Genome-phenome Archive at the European Bioinformatics Institute (https://ega-archive.org) with accession number EGAS00001005567.

Statistical analyses

Unless otherwise stated above, statistical analyses were performed using Prism (v9.0.3, GraphPad, CA) and a summary of analyses is included in the Supplementary material.

Results

Patients and samples

We identified 54 patients with early-stage (defined as stage IA, IB, IC and IIA using the FIGO classification at the time of diagnosis) ovarian high grade serous carcinoma from the pathology archives of three large UK gynaecological cancer centres (Imperial College Healthcare, University College London and Barts Health NHS Trusts). A summary of the workflow is shown in **Fig. 1** and clinical details are given in **Table S1**. Following pathology review, 21 samples from 13 patients were excluded, whilst two samples from one patient failed DNA extraction. Additionally, we identified a further cohort of three early-stage patients recruited into the BriTROC-1 study (15), giving a total early-stage population of 43. The comparison late-stage cohort consisted of 52 patients with stage IIIC/IV disease recruited into the BriTROC-1 study (**Table S1**). The early-stage patients were diagnosed more recently than the late stage (early, median 68 months, range 24-177, prior to analysis; late median 101 months, range 60-179; p=0.0009 **Fig. S1**). The median age at diagnosis for early and late-stage cohorts did not differ significantly (early 61.3 years, range 40-84; late 62.3 years, range 34-76) but overall survival was, as expected, significantly longer in the early-stage cohort than for the late-stage (Hazard Ratio 0.13, 95%CI 0.07-0.26) (**Fig. 2A, B**).

Mutational landscape of early-stage and late-stage cohorts

Using targeted next generation sequencing, we analysed short variants (SNV, indels) in both cohorts (**Table S2**). Mutations in *TP53* were near-universal (100% late-stage patients [52/52]. 89% [34/38] early-stage patients. **Fig. 2C, Fig. S2**). One early-stage case, patient ES_0007, contained two *TP53* missense mutations (L139V, Y163C) at mutant allele frequencies of approximately 50% and 25% respectively. Although these could result from the presence of two separate clones, the CN state at the *TP53* locus in this case was neutral (log_2 ratio shift = -0.056), suggesting that these might be bi-allelic mutations, as we have previously identified in ovarian squamous cell carcinomas arising from mature cystic teratoma (27). The four early-stage samples in which *TP53* mutations were not identified underwent pathology re-review; all were still considered to be HGSC. Two had copy number profiles consistent with HGSC, whilst two showed no CN abnormalities, suggesting very low tumour cellularity (**Fig. S3**). Overall, the frequency of four key *TP53* hotspot mutations (R175, R273, R248, Y220) was significantly greater in the early-stage cohort compared to late (**Fig. 2D, E.** Fisher's exact test; p=0.0029), but there was no difference in the rates of mutations in the other analysed genes (**Fig. 2C**). Specifically, rates of pathological mutations in *BRCA1* and *BRCA2* did not differ significantly between early- and late-stage patients: *BRCA1* 11% (4/38) early vs 17% (9/52) late; *BRCA2* 0% (0/38) early vs 2.0% (1/52) late.

Focal amplifications and deletions in early-stage and late-stage cohorts

We used shallow whole genome sequencing to analyse genome-wide absolute copy number. There was no statistically significant difference in purity between the cohorts (**Fig. 3A**), but median ploidy was significantly greater in late-stage samples compared to early (**Fig. 3B**. Median early 1.9; median late 3.0; p<0.0001, Mann-Whitney test). Global copy number gains/losses are shown in **Fig. 3C** - there were generally more gains and amplifications in late-stage samples, and regions of CN loss in the early-stage samples, in keeping with the differential ploidy between the cohorts. Although there were several regions of differential gain in the late-stage cohort (e.g. chromosome 4, 6, 9, 11,12) and losses in the early-stage cohort (e.g. chromosome 4, 9, 12, 17), we found no significant differences in rates of focal amplification and deletion of 17 genes that are frequently altered in HGSC (7,12) (**Fig. 3D, Fig. S4, Table S3**). The commonest amplifications were in *MYC* (25% in the early-stage and 19% in the late-stage) and *MECOM* (20% in the early-stage and 14% in the late-stage).

Copy number signatures in early-stage and late-stage cohorts

Next, we assessed the distribution of the six specific CN features - segment length, segment copy number, number of breakpoints per chromosome arm, number of breakpoints per 10Mb, copy number change point and length of chains of oscillating copy number (**Fig. S5, 6**) - and used these to generate CN signature exposures for both cohorts (**Fig. 4A, B**). We used a fixed-effects (Bernoulli) analysis to model the presence or absence of signatures and a fixed-effects multivariate normal distribution model, based on isometric log ratio (ILR)-transformation, to compare the two cohorts (see Supplementary Methods). Overall, in the ILR analysis, there was a highly significant difference between cohorts (generalised Wald test; p=7.234e−10), with greater signature 3 exposure in the early-stage cohort and more signature 4 in the late-stage cohort (**Fig. 4C**). In keeping with this, the presence of signature 3 and the absence of signature 4 were both associated with improved overall survival across all patients (**Fig. 4D**).

We then visualised CN signature exposures using simplex plots (**Fig. 4E**) comparing signature 3 (S3), signature 4 (S4) and all other signatures (1-S3-S4). In the early cohort, the sample observations (red dots) cluster towards the top of the right side of the simplex, in keeping with low or zero signature 4 exposure. For the late group, although some of the observations remain in the same place, many are located towards the left of the simplex, indicating that they have non-zero exposure to signature 4, with a relative decrease in the amount of signature 3. The relative contribution of the other signatures does not change between early and late cohorts - the distance from the observations to the top apex of the plot remains similar. Together, this suggests that, overall, signature 3 decreases in relative intensity and signature 4 increases in relative intensity in the late-stage samples, whilst the rest of the signatures remain approximately constant. However, the observations that remain towards the top right of the simplex suggest that a subset of late-

stage samples have genomic features more reminiscent of early-stage. In keeping with this, the presence of signature 3 remained significantly associated with improved overall survival in the late-stage cases, whilst there was a trend for poorer survival with any exposure to signature 4 (**Fig. S7**).

We then performed unsupervised hierarchical clustering of the copy number signature exposures across both cohorts and identified three clusters (**Fig. 5A, Fig. S8**). Cluster 1 had the highest exposure to CN signature 3, cluster 2 was dominated by genomes with high signature 1 exposure, whilst cluster 3 showed highest signature 4 exposure (**Fig. 5B**). There was a significant difference in sample distribution between clusters ($p<0.0001$, Chi-squared) with nearly all early-stage samples in clusters 1 and 2, whilst late-stage samples were spread across all three clusters, with the majority in cluster 3 (**Fig. 5C**). In the 12 patients with >1 sample, samples clustered within the same cluster group in 11 cases (**Fig. S8**). The clusters were prognostic, with a significant trend for reduced survival across the clusters (**Fig. 5D**), which remained significant by Cox proportional hazards (**Fig. 5E**).

We next quantified ploidy by cluster and found highly significant differences across the whole cohort (**Fig. 6A**) with median ploidy of 1.9, 2.6 and 3.3 in clusters 1, 2 and 3 respectively. High ploidy can reflect whole genome duplication (WGD), which is frequent in HGSC (28). However, identifying WGD definitively requires assessment of allele-specific copy number, which is not possible using sWGS or the limited next generation sequencing panel that we utilised. Nonetheless, we previously identified a strong statistical association between CN signature 4 exposure and WGD (10): here, we found a very strong correlation (Spearman rho = 0.73, $p<0.0001$) between ploidy and signature 4 exposure across both cohorts (**Fig. 6B**) implying that the high ploidy may indeed be driven by WGD. In addition, WGD is likely to generate high ploidy states across the entire genome and we found a significantly greater proportion of segments with ploidy >3.0 in clusters 2 and 3 compared to cluster 1 (**Fig. 6C**) as well as in late-stage compared to early-stage samples (**Fig. S9**). Analysis of copy number changepoint (the change in CN state between adjacent segments) also indicated that clusters 2 and 3 had a significantly greater fraction of segments with changepoint of at least +2 than cluster 1, again in keeping with WGD (29) (**Fig. 6D**).

Finally, it was previously shown that median ploidy for cancers with and without WGD were 3.3 and 2.1 respectively (28). Using these delineators, we analysed overall survival of the whole cohort (**Fig. 6E**) and of late-stage patients (**Fig. S10**) and identified significant differences in both analyses.

Discussion

Comparing genomic profiles between cancers with high levels of chromosomal instability is challenging. We have previously developed copy number (CN) signatures that deconvolute copy number features from whole genome analysis to identify the underlying mutational processes that shape the genome. In this manuscript, we have used novel methods to compare CN signature exposures and reveal significant genomic differences between early- and late-stage HGSC. This is clinically important because the large majority of patients with HGSC have advanced disease at the time of diagnosis, reflecting the ease with which HGSC disseminates from the fallopian tube throughout the peritoneal cavity. An important question is whether early-stage HGSC is identified fortuitously through early development of symptoms or whether they have discrete characteristics that reduce the likelihood of peritoneal dissemination, and whether late-stage samples have acquired evolutionary fitness that facilitates metastatic spread.

As expected, we found that *TP53* mutations were near-universal, although four early-stage samples were *TP53* wild type, possibly due to a combination of low tumour cellularity and poor DNA quality from FFPE preservation. A previous study of 16 early-stage HGSC cases (30) also identified that *TP53* mutations were very frequent but not universal, and two of our *TP53* wild type samples certainly had CN profiles that were highly consistent with HGSC. Together, these data suggest that a small proportion of early stage HGSC cases may be truly *TP53* wild type. The rate of missense *TP53* mutations here was higher (75%) than in previous HGSC cohorts (7) and hotspot mutations (defined here as mutation at the four most commonly mutated codons, R248, R273, R175 and Y220) were more prevalent in our early-stage cohort than late-stage. However, a recent analysis of nearly 800 HGSC cases (78 stage I/II, 709 stage III/IV) has suggested no overall difference in mutation type between early and late samples (31). Crucially, we found no differences in rates of *BRCA1/2* mutations between early and late cohorts. The absence of germline DNA from most of our early-stage cohort meant that we were unable to verify the germline mutation status but our *BRCA1/2* mutation results are broadly in line with previous cohorts (32) and reflect the fact that these samples were obtained from routine practice rather than risk reducing surgery. When examining global copy number change, there were no regions uniquely lost or gained in either cohort, suggesting that the process of dissemination is not driven by specific amplifications or deletions, whilst we also found no significant difference in copy number of 17 key genes. Together, the SNV/indel and focal CNA data corroborate findings by Köbel et al that early and late-stage HGSC appear identical by immunohistochemistry (33). The previous WGS analysis of early stage HGSC (30) also identified high levels of genomic instability with few, if any, recurrent focal differences between early and late-stage HGSC, and no unique mutation or focal CN alterations in the early-stage patients.

Our most striking observation was the difference in overall ploidy between early and late cases, which was further reflected in CN signature exposures. Comparison of CN signatures between samples and cohorts is complex because signature exposures are compositional (i.e. they sum to 1 in each sample) and are thus not independent variables: any decrease in one signature will, by definition, be mirrored by an increase in at least one other. In addition, classical statistical methods for analysing compositional data are poor at dealing with zero proportions and many samples have zero exposure to at least one signature. However, using isometric-log ratio analysis of non-zero signature exposures, we found a significant difference between the cohorts overall, driven by signatures 3 (higher in early-stage) and 4 (higher in late-stage). The features that define signature 4 are high segment copy numbers and high copy number changes, both of which are greater in the late cohort. The simplex analysis indicates that the late-stage genomes overall have increased signature 4, although a proportion remains 'early-like' with prominent CN signature 3. The unsupervised hierarchical clustering identified three patterns in the signatures. Clusters 1 and 2 contained most of the early-stage samples, whilst the late-stage samples were divided between the clusters. However, cluster 3 contained almost exclusively late-stage samples and was associated both with higher ploidy and worse survival.

The higher copy number, greater overall ploidy and greater signature 4 exposure all suggest that a proportion of the late-stage tumours had undergone whole genome duplication (WGD). WGD has been described in many solid malignancies (28) and is generally associated with poor prognosis (34). It is thought to arise from aberrant cell division (35) and potentially may mitigate the effects of mutations that would otherwise be deleterious, thus preventing cancer cell attrition (36). In pan-cancer analysis, HGSC has been shown to have one of the highest rates of WGD at approximately 40% (28).

Definitive demonstration of WGD is challenging and requires assessment of both ploidy and extent of LOH (29), which is not possible using sWGS and the targeted capture sequencing panel employed here. In practical terms, the analysis of early stage HGSC requires use of FFPE material that precludes deep WGS analysis. This is an important limitation of our findings. However, our original analysis using deep whole genome sequencing analysis of HGSC specimens showed that CN signature 4 was significantly associated with WGD (10) and here we observed a strong correlation between signature 4 and ploidy, suggesting that the high ploidy samples in cluster 3 may have undergone WGD. We also confirmed other features suggestive of WGD, including that cluster 3 samples were more likely to have high CN across the whole genome and high CN changepoint between segments. Lastly, the median ploidy of early and late cohorts was similar to pan-cancer analyses showing that median ploidy of WGD tumours is 3.3, compared to 2.1 in those lacking WGD (28). The commonest genomic correlate of WGD in previous analyses was mutation in *TP53*, which usually precedes the duplication, as well as *CCNE1* amplification and loss of *RB1*

(28). Although the rates of *CCNE1* amplification here did not differ significantly between our early and late-stage cohorts, our data support the idea that high ploidy is associated with advanced HGSC and poorer prognosis, a concept first explored over twenty years ago (37). HGSC has profound levels of segregation error during cell division, an essential precursor of aneuploidy and WGD (38), and recent data suggest that WGD can emerge in *hTERT*-immortalised human fallopian tube epithelial cells with loss of wild-type p53 function in the presence of *BRCA1* mutation and *MYC* over-expression, but not with *TP53* mutation alone (39).

Although this cohort represents one of the largest collections of early-stage HGSC samples that have been characterised with WGS, this project has potential shortcomings. Our cohort is small, reflecting the rarity of this patient population, which limits our statistical power, in particular the ability to compare early-stage cases that relapsed to those that did not. In addition, the samples were identified retrospectively from pathology archives; thus, the extent of surgery was not defined, and chemotherapy regime given was at the discretion of the treating oncologists. The early-stage cohort was diagnosed more recently, on average, than the late stage, although this difference is unlikely to have influenced clinical management, and nearly 80% were alive 10 years following diagnosis, in keeping with data from randomised clinical trials (5). This strongly suggests that our cohort did not contain large numbers of understaged patients with occult stage III disease. The samples were all formalin-fixed, paraffin-embedded (FFPE) and analysed up to 15 years following diagnosis, and it was not possible to estimate absolute copy number, and hence CN signatures, on several of the samples. Our previously developed methods allow reliable analysis of genome-wide changes in fixed material (40), but demand strict quality control criteria to ensure robust copy number determination. Our estimation of the number of reads required (see Methods) was clearly insufficient in some cases, especially those with low tumour cellularity. If sWGS is to be developed for use in clinical trials, it will be imperative to establish robust criteria for sequencing depth, especially in low cellularity samples or small core biopsies. We also did not perform whole exome sequencing or deep WGS, so are unable to comment upon small variants (SNV, indel) beyond our targeted panel nor on larger scale rearrangements (12,41).

The critical outstanding question is whether the processes that generate high ploidy are the primary drivers of rapid dissemination in HGSC, or whether high ploidy/WGD are simply time-related markers of evolutionary fitness and are thus more likely to be observed in late-stage than early-stage disease. The absence of an age difference between our two cohorts may suggest that the genomic differences are not time-related, consistent with the finding that WGD is an early event in colorectal (35) and non-small cell lung carcinomas (36). In addition, there was close clustering of samples in patients with >1 sample, again suggesting that WGD does not appear as a late event. However, definitive assessments of true rates of WGD in early HGSC compared to advanced disease will require WGS analysis of prospectively collected samples and detailed

comparison between fallopian tube primary site and multi-site examination of large cohorts of disseminated late-stage HGSC as well as *in vitro* models.

In summary, our results indicate that early and late-stage HGSC are similar but also that there may be critical differences, potentially resulting from the appearance of whole genome duplication in a subset of late-stage disease, which is associated with poor outcome. However, our data, reinforced by the striking difference in overall survival in our cohorts, highlight once again the importance of improving strategies that will allow early detection of HGSC.

Acknowledgements

This work was funded by an Imperial/China Scholarship Council scholarship to ZC, the NIHR Imperial Biomedical Research Centre (grant number P77646), Ovarian Cancer Action (grant number 006), the Wellcome Trust (grant number RG92770) and Cancer Research UK (grant numbers A15973, A15601, A18072, A17197, A19274 and A19694). Imperial samples were provided by the Imperial College Healthcare Tissue Bank, which is supported by the NIHR Imperial Biomedical Research Centre. Other infrastructure support was provided by Experimental Cancer Medicine Centres at participating sites and the Cancer Research UK Imperial Centre.

Author contributions

Study design: IAMcN, FM, JDB
Sample acquisition: ML, LT, JK, DE, BK, NS, JMcD
Pathological assessment: BK, NS, CS, JMcD, DE
Data acquisition: ZC, HM, PS, GG, DE
Data analysis: ZC, HM, PS, LM, TG, TB, AP, GM
Manuscript preparation: ZC, IAMcN
All authors reviewed the manuscript before submission

References

1. Perren TJ, Swart AM, Pfisterer J, Ledermann JA, Pujade-Lauraine E, Kristensen G*, et al.* A phase 3 trial of bevacizumab in ovarian cancer. *N Engl J Med* 2011;**365**(26):2484-96 doi 10.1056/NEJMoa1103799 [doi].
2. Burger RA, Brady MF, Bookman MA, Fleming GF, Monk BJ, Huang H*, et al.* Incorporation of bevacizumab in the primary treatment of ovarian cancer. *N Engl J Med* 2011;**365**(26):2473-83 doi 10.1056/NEJMoa1104390 [doi].
3. González-Martín A, Pothuri B, Vergote I, DePont Christensen R, Graybill W, Mirza MR*, et al.* Niraparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. *N Engl J Med* 2019;**381**(25):2391-402 doi 10.1056/NEJMoa1910962.
4. Ray-Coquard I, Pautier P, Pignata S, Perol D, Gonzalez-Martin A, Berger R*, et al.* Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. *N Engl J Med* 2019;**381**(25):2416-28 doi 10.1056/NEJMoa1911361.
5. Collinson F, Qian W, Fossati R, Lissoni A, Williams C, Parmar M*, et al.* Optimal treatment of early-stage ovarian cancer. *Ann Oncol* 2014;**25**(6):1165-71 doi 10.1093/annonc/mdu116.
6. Ahmed AA, Etemadmoghadam D, Temple J, Lynch AG, Riad M, Sharma R*, et al.* Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *J Pathol* 2010;**221**(1):49-56 doi 10.1002/path.2696 [doi].
7. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**(7353):609-15.
8. Lee Y, Miron A, Drapkin R, Nucci MR, Medeiros F, Saleemuddin A*, et al.* A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *J Pathol* 2007;**211**(1):26-35 doi 10.1002/path.2091.
9. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* 2013;**45**(10):1127-33 doi 10.1038/ng.2762.
10. Macintyre G, Goranova T, De Silva D, Ennis D, Piskorz AM, Eldridge M*, et al.* Copy-number signatures and mutational processes in ovarian carcinoma. *Nat Genet* 2018;**50**(9):1262-70 doi 10.1038/s41588-018-0179-8.
11. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S*, et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 2008;**14**(16):5198-208.
12. Patch A-M, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S*, et al.* Whole–genome characterization of chemoresistant ovarian cancer. *Nature* 2015;**521**(7553):489-94 doi 10.1038/nature14410
13. Eckert MA, Pan S, Hernandez KM, Loth RM, Andrade J, Volchenboum SL*, et al.* Genomics of Ovarian Cancer Progression Reveals Diverse Metastatic Trajectories Including Intraepithelial Metastasis to the Fallopian Tube. *Cancer Discov* 2016;**6**(12):1342-51 doi 10.1158/2159-8290.cd-16-0607.
14. Labidi-Galy SI, Papp E, Hallberg D, Niknafs N, Adleff V, Noe M*, et al.* High grade serous ovarian carcinomas originate in the fallopian tube. *Nat Commun* 2017;**8**(1):1093 doi 10.1038/s41467-017-00962-1.
15. Goranova T, Ennis D, Piskorz AM, Macintyre G, Lewsley LA, Stobo J*, et al.* Safety and utility of image-guided research biopsies in relapsed high-grade serous ovarian carcinoma-experience of the BriTROC consortium. *Br J Cancer* 2017;**116**(10):1294-301 doi 10.1038/bjc.2017.86.
16. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 2009;**25**(14):1754-60 doi 10.1093/bioinformatics/btp324.
17. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellstrom-Lindberg E, Jansen JH*, et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific reports* 2017;**7**:43169 doi 10.1038/srep43169.
18. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 2019:861054 doi 10.1101/861054.
19. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L*, et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 2012;**22**(3):568-76 doi 10.1101/gr.129684.111.

20. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)* 2012;**28**(14):1811-7 doi 10.1093/bioinformatics/bts271.
21. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA*, et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 2018:201178 doi 10.1101/201178.
22. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A*, et al.* The Ensembl Variant Effect Predictor. *Genome biology* 2016;**17**(1):122 doi 10.1186/s13059-016-0974-4.
23. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF*, et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome research* 2014;**24**(12):2022-32 doi 10.1101/gr.175141.114.
24. Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software; Vol 1, Issue 5 (2016)* 2016 doi 10.18637/jss.v070.i05.
25. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* 2014;**61**(6):1-36 doi 10.18637/jss.v061.i06.
26. Therneau T. 2020 A Package for Survival Analysis in R. < https://CRAN.R-project.org/package=survival.>.
27. Cooke SL, Ennis D, Evers L, Dowson S, Chan MY, Paul J*, et al.* The driver mutational landscape of ovarian squamous cell carcinomas arising in mature cystic teratoma. *Clin Cancer Res* 2017;**34**(24):7633-40 doi 10.1158/1078-0432.ccr-17-1789.
28. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J*, et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* 2018;**50**(8):1189-95 doi 10.1038/s41588-018-0165-1.
29. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG*, et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* 2021;**184**(8):2239-54.e39 doi 10.1016/j.cell.2021.03.009.
30. Chien J, Sicotte H, Fan JB, Humphray S, Cunningham JM, Kalli KR*, et al.* TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. *Nucleic Acids Res* 2015;**43**(14):6945-58 doi 10.1093/nar/gkv111.
31. Tuna M, Ju Z, Yoshihara K, Amos CI, Tanyi JL, Mills GB. Clinical relevance of TP53 hotspot mutations in high-grade serous ovarian cancers. *Br J Cancer* 2020;**122**(3):405-12 doi 10.1038/s41416-019-0654-8.
32. Rust K, Spiliopoulou P, Tang CY, Bell C, Stirling D, Phang THF*, et al.* Routine germline BRCA1 and BRCA2 testing in ovarian carcinoma patients: analysis of the Scottish real life experience. *Bjog* 2018;**125**(11):1451-8 doi 10.1111/1471-0528.15171.
33. Köbel M, Kalloger SE, Boyd N, McKinney S, Mehl E, Palmer C*, et al.* Ovarian Carcinoma Subtypes Are Different Diseases: Implications for Biomarker Studies. *PLoS medicine* 2008;**5**(12):e232.
34. Storchova Z, Pellman D. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol* 2004;**5**(1):45-54 doi 10.1038/nrm1276.
35. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Grönroos E, Endesfelder D*, et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov* 2014;**4**(2):175-85 doi 10.1158/2159-8290.Cd-13-0285.
36. López S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M*, et al.* Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat Genet* 2020;**52**(3):283-93 doi 10.1038/s41588-020-0584-7.
37. Tropé C, Kaern J, Hogberg T, Abeler V, Hagen B, Kristensen G*, et al.* Randomized study on adjuvant chemotherapy in stage I high-risk ovarian cancer with evaluation of DNA-ploidy as prognostic instrument. *Ann Oncol* 2000;**11**(3):281-8 doi 10.1023/a:1008399414923.
38. Nelson L, Tighe A, Golder A, Littler S, Bakker B, Moralli D*, et al.* A living biobank of ovarian cancer ex vivo models reveals profound mitotic heterogeneity. *Nat Commun* 2020;**11**(1):822 doi 10.1038/s41467-020-14551-2.

39. Bronder D, Wangsa D, Zong D, Meyer TJ, Wardenaar R, Minshall P*, et al.* TP53 loss initiates chromosomal instability in high-grade serous ovarian cancer. *bioRxiv* 2021:2021.03.12.435079 doi 10.1101/2021.03.12.435079.
40. Piskorz AM, Ennis D, Macintyre G, Goranova TE, Eldridge M, Segui-Gracia N*, et al.* Methanol-based fixation is superior to buffered formalin for next-generation sequencing of DNA from clinical cancer samples. *Ann Oncol* 2016;**27**(3):532-9 doi 10.1093/annonc/mdv613.
41. Ewing A, Meynert A, Churchman M, Grimes G, Hollis RL, Herrington CS*, et al.* Structural variants at the BRCA1/2 loci are a common source of homologous repair deficiency in high grade serous ovarian carcinoma. *Clin Cancer Res* 2021;**27**(11):3201-14 doi 10.1158/1078-0432.Ccr-20-4068.

Figures Legends

Figure 1. REMARK diagram for early-stage and late-stage cohorts

Figure 2. Clinical features and mutational landscape of early-stage and late-stage cohorts.

(A) Diagnosis age. Median 61.3 year (early stage), 62.3 years (late stage). p=NS
(B) Overall survival. Median 60.3 months for late stage, and not reached for early stage. Log-rank Hazard Ratio 0.13 (95%CI 0.07-0.26), p<0.0001 (Log-rank).
(C) Short variants (SNV and indels) for each patient in early-stage and late-stage cohorts. The upper plot shows the number of mutations in each tumour sample.
(D) Gene mutation mapper plot of *TP53* in early-stage cohort and (E) late-stage cohort. Key hotspot residues are marked. The commonest residue mutations in each cohort are marked in red

Figure 3. Focal gene amplifications and deletions in early-stage and late-stage cohorts.

(A) Purity comparison of early-stage and late-stage cohorts.
(B) Ploidy comparison of early-stage and late-stage cohorts; Mann Whitney test.
(C) Global copy number amplifications, gains and losses in early-stage and late-stage cohorts.
(D) Estimation of focal amplifications and deletions in 17 genes of interest, determined by sWGS. The upper plot shows the number of amplifications and deletions in each tumour sample.

Figure 4. Copy number signatures in early-stage and late-stage cohorts.

(A) Copy number signature exposures in early-stage and late-stage patients. Note that signature exposures sum to 1 in each sample. Bars above signatures indicate adjacent samples derived from the same patient.
(B) Mean signature exposure proportions across the early-stage and late-stage cohorts.
(C) Comparison of signature exposures across early-stage and late-stage cohorts; Wald test.
(D) Overall survival of combined early and late-stage cohorts by zero vs non-zero exposures to copy number signature 3 (left) and 4 (right). Log-rank (Mantel-Cox) analysis.
(E) Simplex plots representing exposures for CN signature 3 (right axis), signature 4 (bottom axis) and the rest of the signatures (1 - S3 - S4) combined (left axis) in early (left) and late (right) stage

cohorts. Each red dot represents a single sample, and the contours represent the density of observed samples.

Figure 5. Relationship between signature exposures and clinical factors.

(A) Unsupervised hierarchical clustering in combined early-stage and late-stage cohorts.
(B) Distributions of copy number signature exposures in three clusters
(C) Early and late-stage samples by cluster; Chi-squared test.
(D) Overall survival by cluster; Log-rank for trend
(E) Forrest plot of hazard ratio estimates on overall survival (OS) for clusters. Cox proportional hazards.

Figure 6. Cluster ploidy and whole genome duplication.

(A) Ploidy distribution of late-stage samples in three clusters. Kruskal Wallis test.
(B) Correlation between CN signature 4 exposure and ploidy across both cohorts. Spearman rank correlation.
(C) Fraction of CN segments with absolute copy number ≥3 in three clusters. Kruskal Wallis test.
(D) Copy number changepoint. Graphical depiction of CN changepoint (left); distribution of copy number changepoint ≥+2 in the three clusters. Kruskal Wallis test (centre); density distribution (right).
(E) Overall survival of combined early and late-stage cohorts by ploidy. Log-rank for trend analysis.

Figure 1

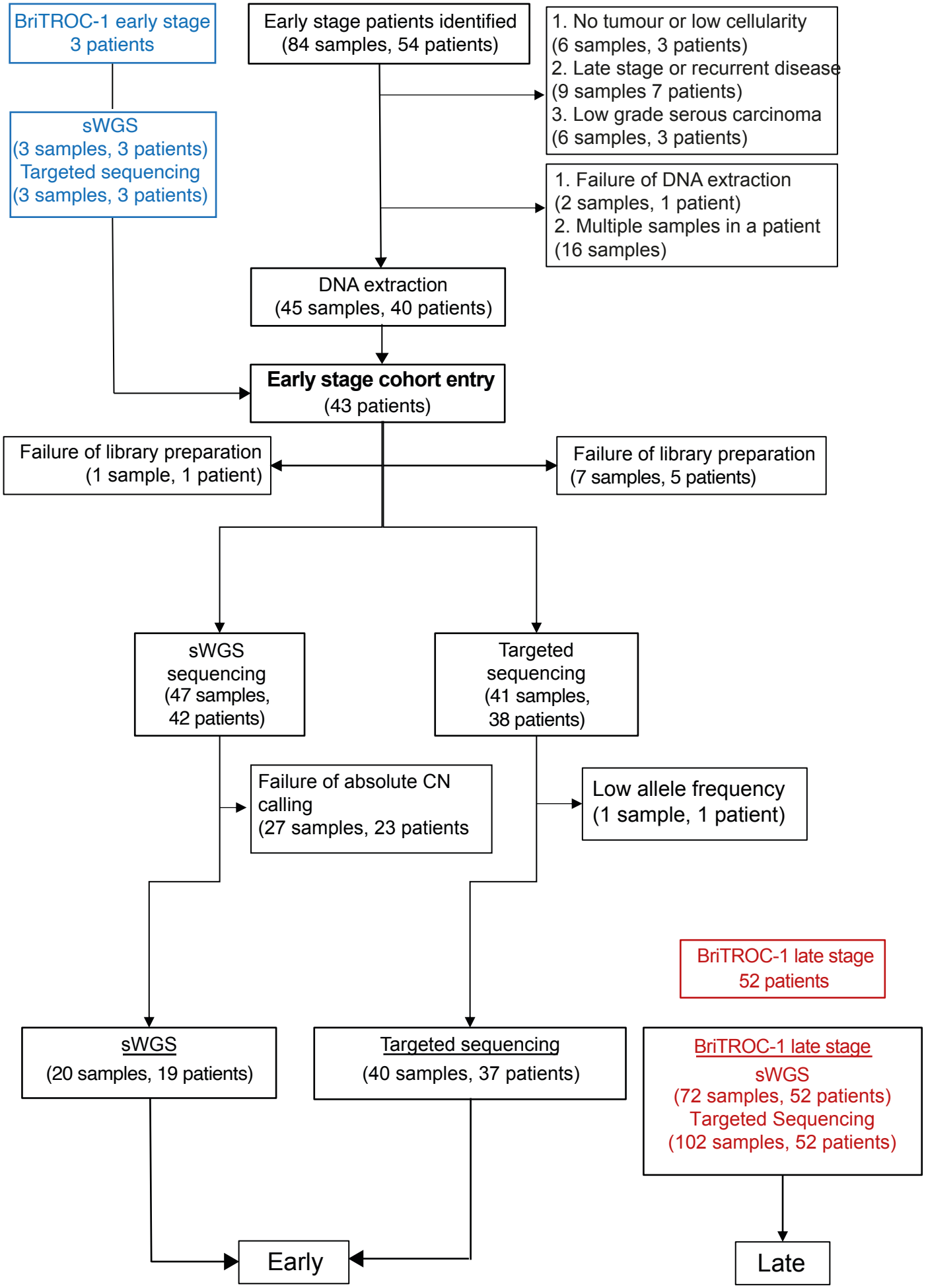

BriTROC-1 early stage
3 patients
Early stage patients identified
(84 samples, 54 patients)
1. No tumour or low cellularity
(6 samples, 3 patients)
2. Late stage or recurrent disease
(9 samples 7 patients)
3. Low grade serous carcinoma
(6 samples, 3 patients)
sWGS
(3 samples, 3 patients)
Targeted sequencing
(3 samples, 3 patients)
1. Failure of DNA extraction
(2 samples, 1 patient)
2. Multiple samples in a patient
(16 samples)
DNA extraction
(45 samples, 40 patients)
Early stage cohort entry
(43 patients)
Failure of library preparation
(1 sample, 1 patient)
Failure of library preparation
(7 samples, 5 patients)
sWGS
sequencing
(47 samples,
42 patients)
Targeted
sequencing
(41 samples,
38 patients)
Failure of absolute CN
calling
(27 samples, 23 patients
Low allele frequency
(1 sample, 1 patient)
BriTROC-1 late stage
52 patients
sWGS
(20 samples, 19 patients)
Targeted sequencing
(40 samples, 37 patients)
BriTROC-1 late stage
sWGS
(72 samples, 52 patients)
Targeted Sequencing
(102 samples, 52 patients)
Early
Late

Figure 2

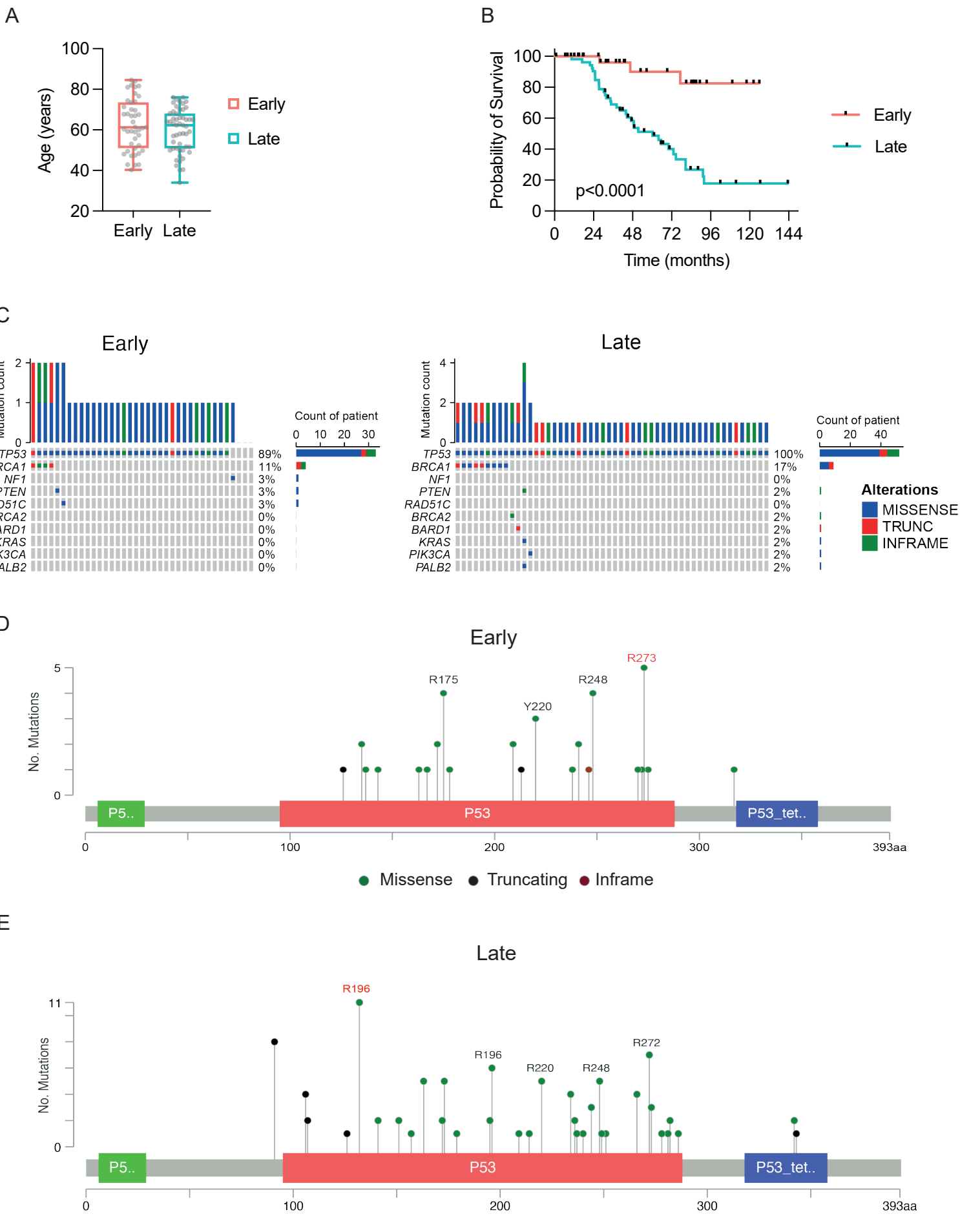
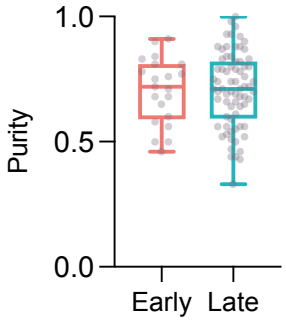
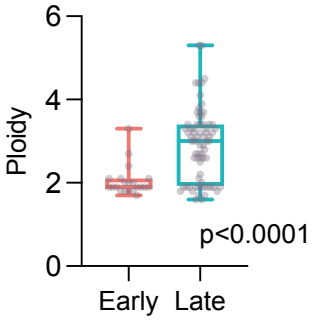

A
Age (years)
100
80
60
40
20
Early
Late
Early
Late
B
Probability of Survival
100
80
60
40
20
0
p<0.0001
0
24
48
72
96
120
144
Time (months)
Early
Late
C
Early
Late
Mutation count
Count of patient
TP53
BRCA1
NF1
PTEN
RAD51C
BRCA2
BARD1
KRAS
PIK3CA
PALB2
89%
11%
3%
3%
3%
0%
0%
0%
0%
0%
100%
17%
0%
2%
0%
2%
2%
2%
2%
2%
Alterations
MISSENSE
TRUNC
INFRAME
D
Early
No. Mutations
R175
Y220
R248
R273
P5..
P53
P53_tet..
0
100
200
300
393aa
Missense
Truncating
Inframe
E
Late
No. Mutations
R196
R196
R220
R248
R272
P5..
P53
P53_tet..
0
100
200
300
393aa

Figure 3

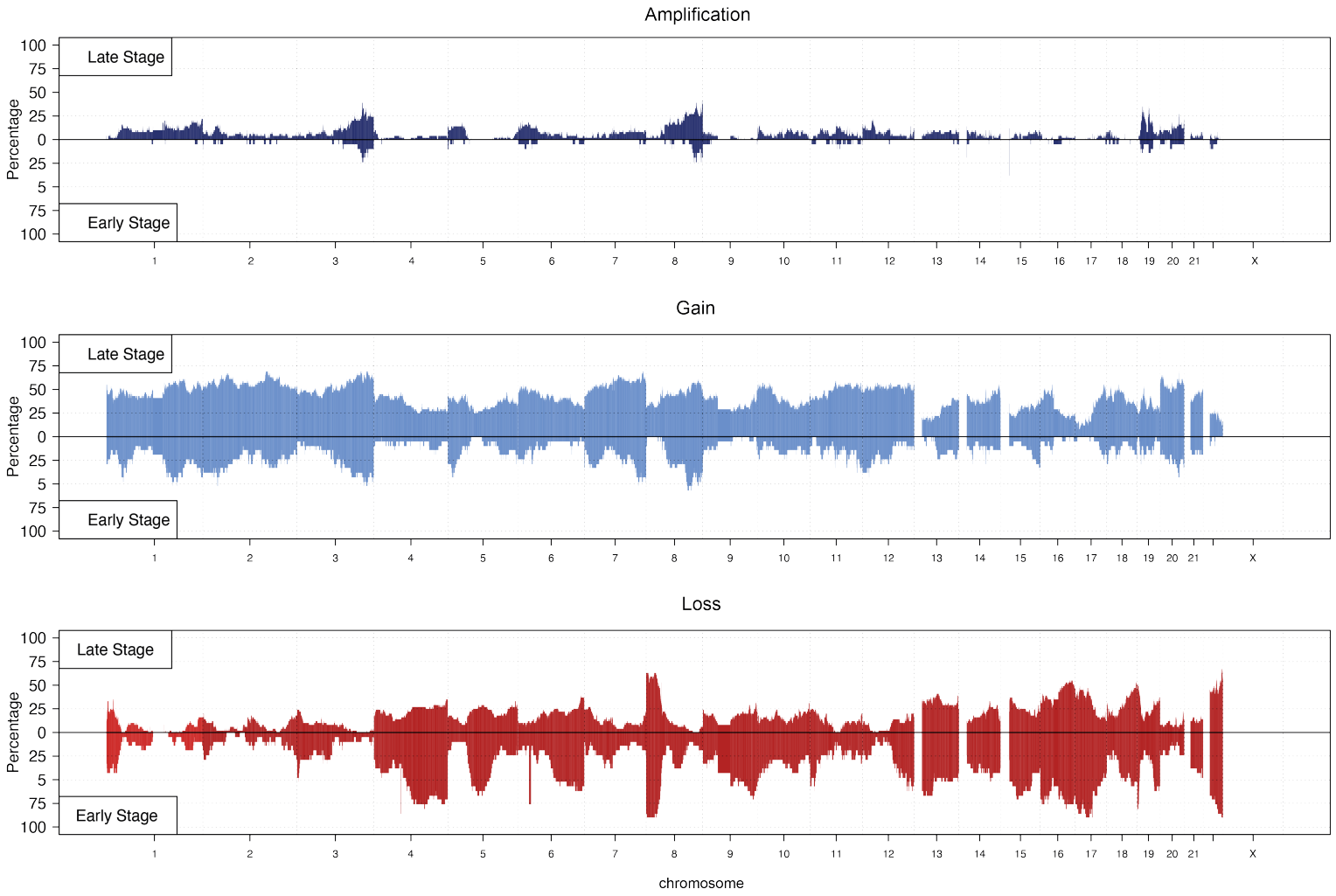
A


1.0
0.5
0.0
Purity
Early
Late

B


6
4
2
0
Ploidy
p<0.0001
Early
Late

C


Amplification
Late Stage
Early Stage
Percentage
100
75
50
25
0
25
5
75
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
X
Gain
Late Stage
Early Stage
Percentage
Loss
Late Stage
Early Stage
Percentage
chromosome

D

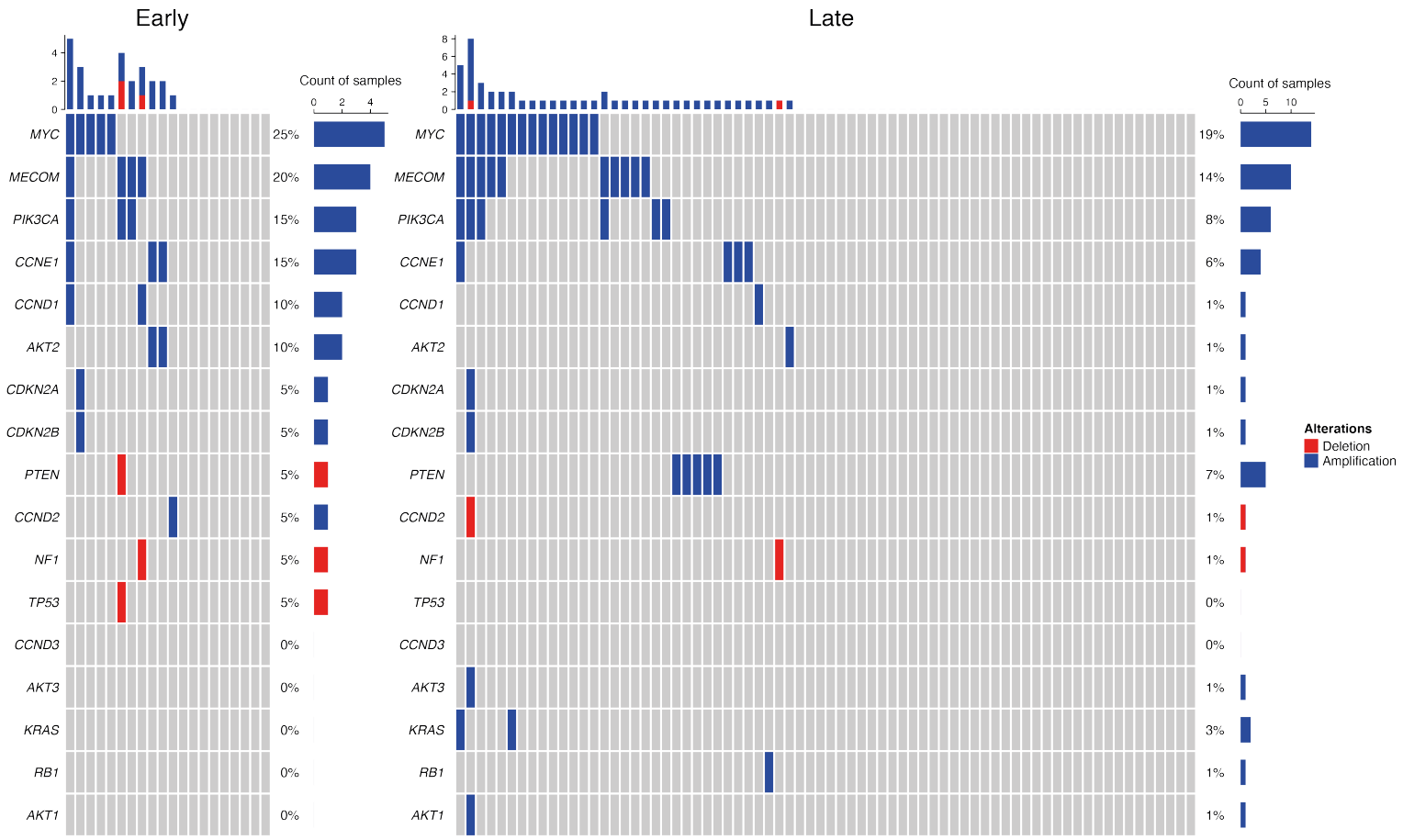
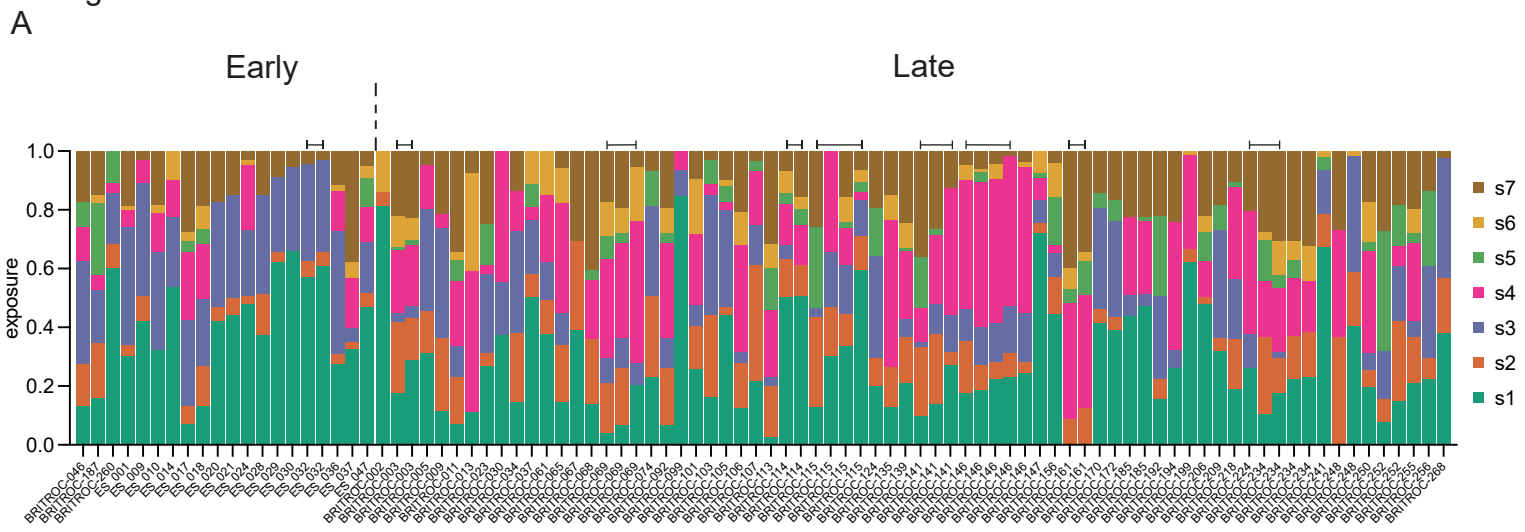
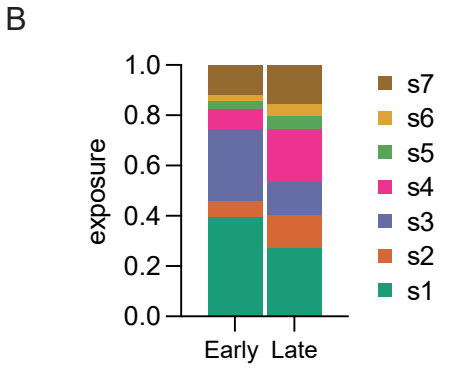
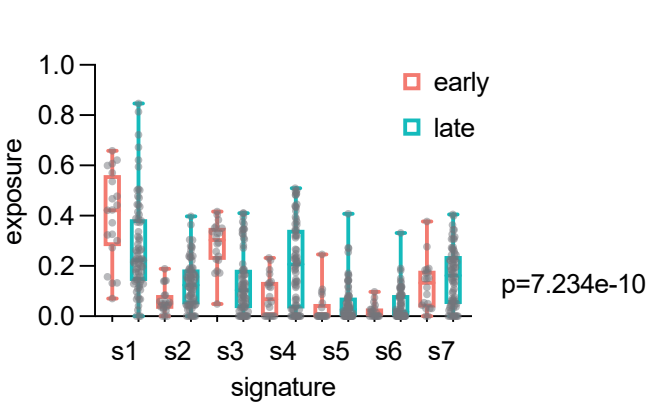
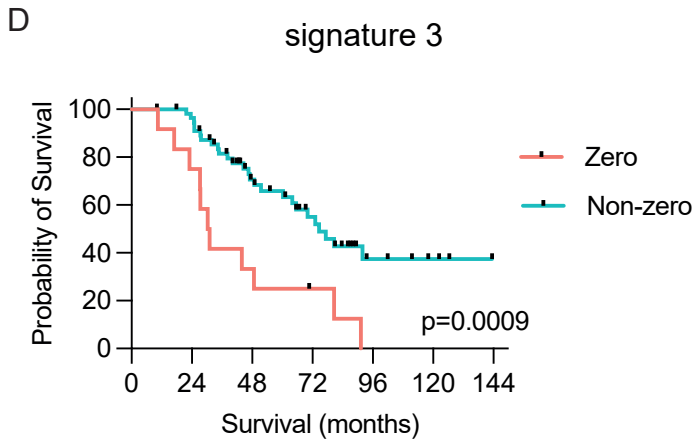

Early
Late
Count of samples
Count of samples
MYC
MECOM
PIK3CA
CCNE1
CCND1
AKT2
CDKN2A
CDKN2B
PTEN
CCND2
NF1
TP53
CCND3
AKT3
KRAS
RB1
AKT1
25%
20%
15%
15%
10%
10%
5%
5%
5%
5%
5%
5%
0%
0%
0%
0%
0%
19%
14%
8%
6%
1%
1%
1%
1%
7%
1%
1%
0%
0%
1%
3%
1%
1%
Alterations
Deletion
Amplification

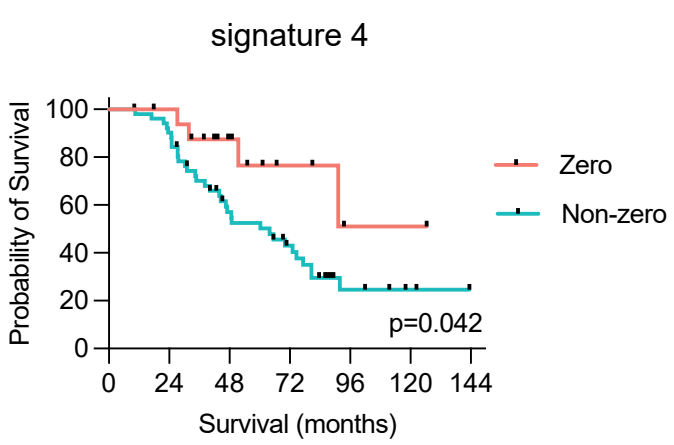
Figure 4


A
Early
Late
exposure
1.0
0.8
0.6
0.4
0.2
0.0
s7
s6
s5
s4
s3
s2
s1


B
exposure
1.0
0.8
0.6
0.4
0.2
0.0
Early
Late
s7
s6
s5
s4
s3
s2
s1


C
exposure
1.0
0.8
0.6
0.4
0.2
0.0
early
late
p=7.234e-10
s1
s2
s3
s4
s5
s6
s7
signature


D
signature 3
Probability of Survival
100
80
60
40
20
0
Zero
Non-zero
p=0.0009
0
24
48
72
96
120
144
Survival (months)


signature 4
Probability of Survival
100
80
60
40
20
0
Zero
Non-zero
p=0.042
0
24
48
72
96
120
144
Survival (months)

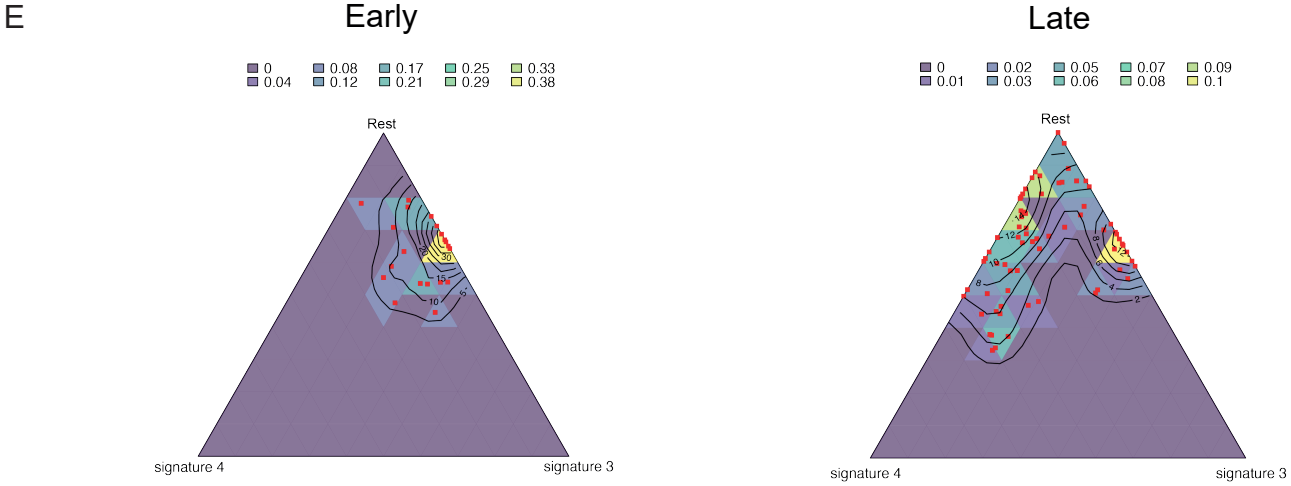

E
Early
0
0.04
0.08
0.12
0.17
0.21
0.25
0.29
0.33
0.38
Rest
signature 4
signature 3
Late
0
0.01
0.02
0.03
0.05
0.06
0.07
0.08
0.09
0.1
Rest
signature 4
signature 3

Figure 5

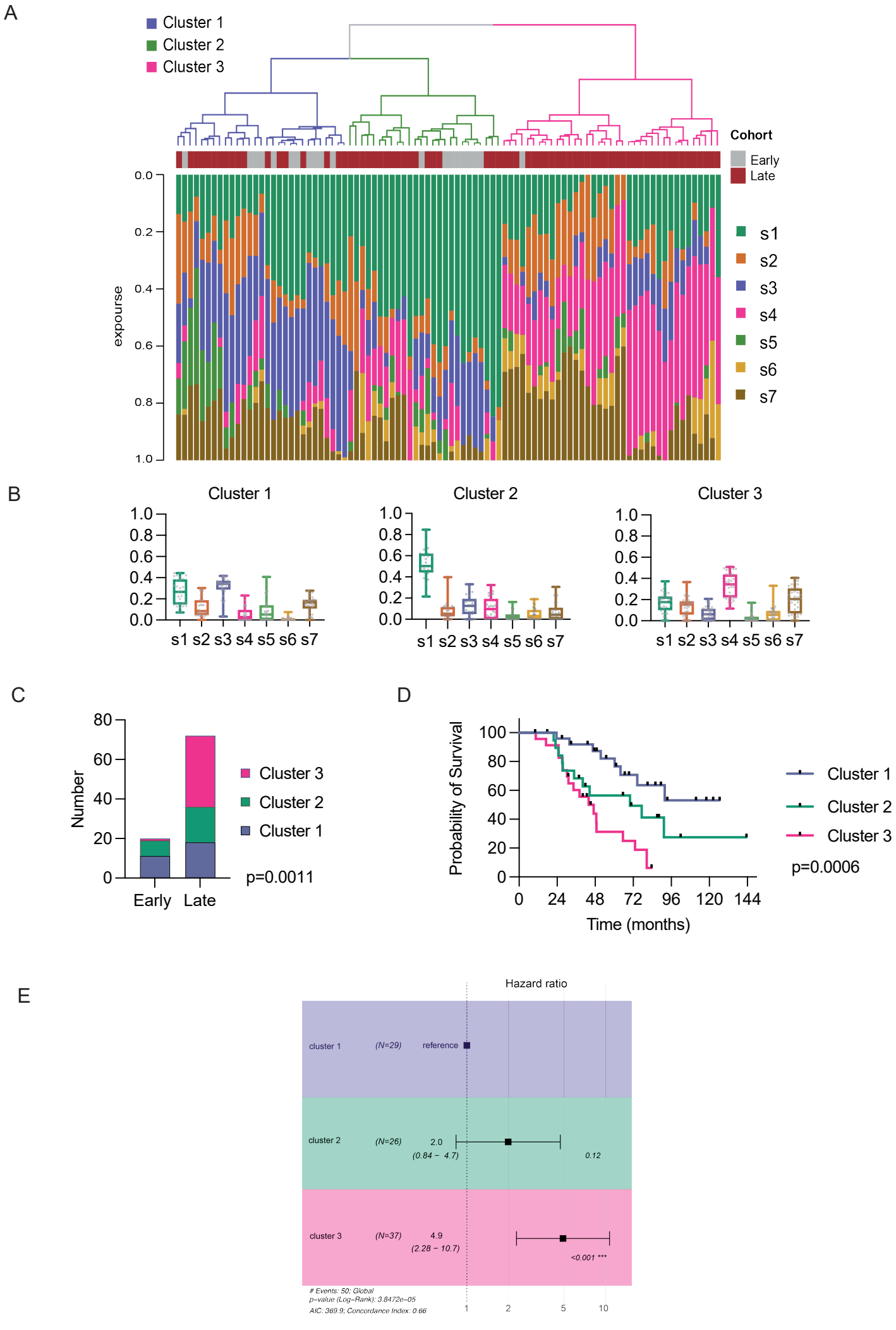
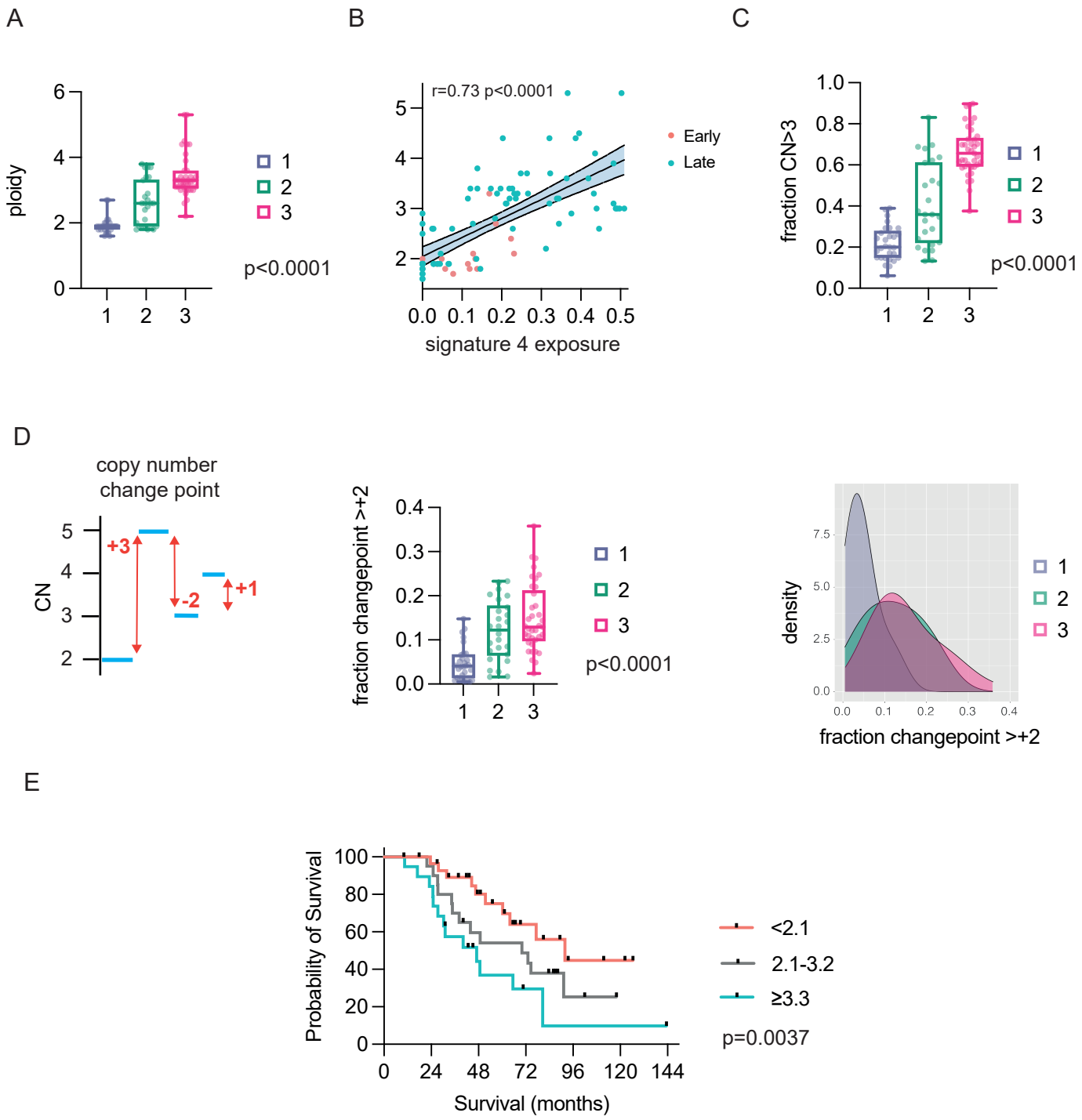

A
Cluster 1
Cluster 2
Cluster 3
Cohort
Early
Late
expourse
0.0
0.2
0.4
0.6
0.8
1.0
s1
s2
s3
s4
s5
s6
s7
B
Cluster 1
Cluster 2
Cluster 3
s1 s2 s3 s4 s5 s6 s7
C
Number
Early
Late
p=0.0011
D
Probability of Survival
Time (months)
p=0.0006
E
Hazard ratio
cluster 1
(N=29)
reference
cluster 2
(N=26)
2.0
(0.84 − 4.7)
0.12
cluster 3
(N=37)
4.9
(2.28 − 10.7)
<0.001 ***
Events: 50; Global
p−value (Log−Rank): 3.8472e−05
AIC: 369.9; Concordance Index: 0.66

Figure 6


A
ploidy
1
2
3
p<0.0001
B
r=0.73 p<0.0001
Early
Late
signature 4 exposure
C
fraction CN>3
p<0.0001
D
copy number
change point
CN
+3
-2
+1
fraction changepoint >+2
p<0.0001
density
fraction changepoint >+2
E
Probability of Survival
Survival (months)
<2.1
2.1-3.2
≥3.3
p=0.0037