Copy-number signatures and mutational processes in ovarian carcinoma

Geoff Macintyre[1 †], Teodora E. Goranova[1 †], Dilrini De Silva[1], Darren Ennis[2], Anna M. Piskorz[1], Matthew Eldridge[1], Daoud Sie[3], Liz-Anne Lewsley[4], Aishah Hanif[4], Cheryl Wilson[4], Suzanne Dowson[2], Rosalind M. Glasspool[5], Michelle Lockley[6,7], Elly Brockbank[8], Ana Montes[9], Axel Walther[10], Sudha Sundar[11], Richard Edmondson[12,13], Geoff D. Hall[14], Andrew Clamp[15], Charlie Gourley[16], Marcia Hall[17], Christina Fotopoulou[18], Hani Gabra[18,19], James Paul[4], Anna Supernat[1], David Millan[20], Aoisha Hoyle[20], Gareth Bryson[20], Craig Nourse[2], Laura Mincarelli[2], Luis Navarro Sanchez[2], Bauke Ylstra[3], Mercedes Jimenez-Linan[21], Luiza Moore[21], Oliver Hofmann[2,22], Florian Markowetz[1,*], Iain A. McNeish[2,5,18,*], James D. Brenton[1,21,23,*,#]

1. Cancer Research UK Cambridge Institute, Cambridge, CB2 0RE, UK
2. Institute of Cancer Sciences, University of Glasgow, G61 1QH, UK
3. VU University Medical Center, Amsterdam 1007 MB, The Netherlands
4. Cancer Research UK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, G12 0YN, UK
5. Beatson West of Scotland Cancer Centre, Glasgow, G12 0YN, UK
6. Barts Cancer Institute, Queen Mary University of London, London, EC1M 6BQ, UK
7. University College London Hospital, London, WC1E 6BD, UK
8. Barts Health NHS Trust, London, E1 1BB
9. Guy's Hospital, London, SE1 9RT, UK
10. Bristol Cancer Institute, Bristol, BS2 8ED, UK
11. City Hospital, Birmingham, B18 7QH, UK
12. Division of Cancer Sciences, Faculty of Biology, Medicine and Health, University of Manchester, St Mary's Hospital, Manchester, UK;
13. Department of Obstetrics and Gynaecology, Manchester Academic Health Science Centre, St Mary's Hospital, Central Manchester NHS Foundation Trust, Manchester Academic Health Science Centre; Level 5, Research, Oxford Road, Manchester, UK;
14. St James's Institute of Oncology, Leeds, LS9 7TF UK
15. The Christie Hospital, Manchester, M20 4BX, UK
16. Nicola Murray Centre for Ovarian Cancer Research, MRC IGMM, University of Edinburgh, Edinburgh, EH4 2XR, UK
17. Mount Vernon Cancer Centre, Northwood, HA6 2RN, UK
18. Division of Cancer and Ovarian Cancer Action Research Centre, Department of Surgery and Cancer, Imperial College, London, W12 0NN, UK
19. Early Clinical Development, IMED Biotech Unit, AstraZeneca, Cambridge, UK
20. Department of Pathology, Queen Elizabeth University Hospital, Glasgow G51 4TF, UK
21. Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK.
22. Centre for Cancer Research, University of Melbourne, VIC 3010 Australia
23. Department of Oncology, University of Cambridge, CB2 0XZ, UK

[†] These authors contributed equally to this work.
* Co-corresponding authors: Florian Markowetz (Florian.Markowetz@cruk.cam.ac.uk), Iain McNeish (i.mcneish@imperial.ac.uk), James Brenton (James.Brenton@cruk.cam.ac.uk)
[#] Lead contact

Abstract

The genomic complexity of profound copy-number aberration has prevented effective molecular stratification of ovarian cancers. To decode this complexity, we derived copy-number signatures from shallow whole genome sequencing of 117 high-grade serous ovarian cancer (HGSOC) cases, which were validated on 527 independent cases. We show that HGSOC comprises a continuum of genomes shaped by multiple mutational processes that result in known patterns of genomic aberration. Copy-number signature exposures at diagnosis predict both overall survival and the probability of platinum-resistant relapse. Measuring signature exposures provides a rational framework to choose combination treatments that target multiple mutational processes.

Introduction

The discrete mutational processes that drive copy-number change in human cancers are not readily identifiable from genome-wide sequence data. This presents a major challenge for the development of precision medicine for cancers that are strongly dominated by copy-number changes, including high-grade serous ovarian (HGSOC), esophageal, non-small-cell lung and triple negative breast cancers[1]. These tumors have low frequency of recurrent oncogenic mutations, few recurrent copy number alterations, and highly complex genomic profiles[2].

HGSOCs are poor prognosis carcinomas with ubiquitous *TP53* mutation[3]. Despite efforts to discover new molecular subtypes and targeted therapies, overall survival has not improved over two decades[4]. Current genomic stratification is limited to defining homologous recombination-deficient (HRD) tumors[5-7] with approximately 20% HGSOC cases having a germline or somatic mutation in *BRCA1/2* with smaller contributions from mutation or epigenetic silencing of other HR genes[8]. Classification using gene expression predominantly reflects the tumor microenvironment and is reliable in only a subset of patients[9-11]. Detailed genomic analysis using whole genome sequencing has shown frequent loss of *RB1, NF1* and *PTEN* by gene breakage events[12] and enrichment of amplification associated fold-back inversions in non-HRD tumors[13]. However, none of these approaches has provided a broad mechanistic understanding of HGSOC, reflecting the challenges of detecting classifiers in extreme genomic complexity.

Recent algorithmic advances have enabled interpretation of complex genomic changes by identifying mutational signatures — genomic patterns that are the imprint of mutagenic processes accumulated over the lifetime of a cancer cell[14]. For example, UV exposure or mismatch repair defects induce distinct, detectable single nucleotide variant (SNV) signatures[14]. The clinical utility of these signatures has recently been demonstrated through a combination of structural variant (SV) and SNV signatures to improve the prediction of HRD[15]. Importantly, these studies show that tumor genomes are shaped by multiple mutational processes and novel computational approaches are needed to identify coexistent signatures. We hypothesized that specific features of copy-number abnormalities could represent the imprints of distinct mutational processes, and developed methods to identify signatures from copy-number features in HGSOC.

Results

Experimental design and data collection

We generated absolute copy number profiles from 253 primary and relapsed HGSOC samples from 132 patients in the BriTROC-1 cohort[16] using low-cost shallow whole-genome sequencing (sWGS; 0.1×) and targeted amplicon sequencing of *TP53* (Supplementary Figure 1). These samples formed the basis of our copy-number signature identification. A subset of 56 of these cases had deep whole-genome sequencing (dWGS) performed for mutation analysis and comparison with sWGS data. Independent data sets for validation included 112 dWGS HGSOC cases from PCAWG[17] and 415 HGSOC cases with SNP array and whole exome sequence from TCGA[8]. Supplementary Figure 1a shows the REMARK diagram for selection of BriTROC-1 patients. Supplementary Figure 1b outlines which samples were used in each analysis across the three cohorts. Clinical data for the BriTROC-1 cohort are summarized in Supplementary Table 1 and Supplementary Figure 2. Detailed information on experimental design is provided in the Life Sciences Reporting Summary.

Identification and validation of copy-number signatures

To identify copy-number (CN) signatures, we computed the genome-wide distributions of six fundamental CN features for each sample: the breakpoint count per 10MB, the copy-number of segments, the difference in CN between adjacent segments, the breakpoint count per chromosome arm, the lengths of oscillating CN segment chains and the size of segments. These features were selected as hallmarks of previously reported genomic aberrations, including breakage-fusion-bridge cycles[18], chromothripsis[19] and tandem duplication[20,21].

We applied mixture modelling to separate the copy-number feature distributions from 91 BriTROC-1 samples with high quality CN profiles into mixtures of Poisson or Gaussian distributions. This resulted in a total of 36 mixture components (Figure 1a). For each sample, the posterior probability of copy-number events arising from these components was computed and summed. These sum-of-posterior vectors were then combined to form a sample-by-component sum-of-posteriors matrix. To identify copy-number signatures, this matrix was subjected to non-negative matrix factorization (NMF)[22], a method previously used for deriving SNV signatures[14].

NMF identified seven CN signatures (Figure 1a), as well as their defining features and exposures in each sample. The optimal number of signatures was chosen using a consensus from 1000 initializations of the algorithm and 1000 random permutations of the data combining four model selection measures (Supplementary Figure 3). We found highly similar component weights for the signatures in the two independent cohorts (PCAWG-OV and TCGA), demonstrating the robustness

129 of both the methodology and the copy-number features (Figure 1b, P<9e-05, median r=0.86.
130 Supplementary Table 2), despite a significant difference in exposures to CN signatures 2, 3, 4 and
131 5 between the cohorts (P<0.05, two-sided Wilcoxon rank sum test, Supplementary Figure 4).

132 Mutational processes underlying copy-number signatures

133 The majority of cases analysed exhibited multiple signature exposures suggesting that HGSOC
134 genomes are shaped by more than one mutational process. As our signature analysis reduced this
135 genomic complexity into its constituent components, we were able to link the individual copy-
136 number signatures to their underlying mutational processes. To do this, we used the component
137 weights identified by NMF to determine which pattern of global or local copy-number change
138 defined each signature. For example, for CN signature 1, the highest weights were observed for
139 components representing low numbers of breakpoints per 10MB, long genomic segments and two
140 breaks occurring per chromosome arm (Figure 2a, Supplementary Figure 5). Two breaks per
141 chromosome arm suggested that the mutational process underlying this signature might be
142 breakage-fusion-bridge (BFB) events[18].

143 To test this hypothesis, we correlated CN signature 1 exposures with mutation data, SNV
144 signatures, and other measures derived from deep WGS and exome sequencing (Figure 2b-e,
145 Supplementary Figures 6, 7, 8 and 9, Supplementary Tables 3, 4, 5, 6, 7 and 8). CN signature 1
146 was anti-correlated with sequencing estimates of telomere length (r=-0.32, P=0.009), consistent
147 with BFB events. In addition, CN signature 1 was positively correlated with amplification-
148 associated fold-back inversion structural variants (r=0.36, P=0.02), which have been strongly
149 implicated in BFB events[23] and have also been associated with inferior survival in HGSOC[13]. CN
150 signature 1 was also enriched in cases with oncogenic RAS signaling, including *NF1* loss and
151 mutated *KRAS* (p=5e-06, Mann-Whitney test), which has previously been shown to induce
152 chromosomal instability as a result of aberrant G2 and mitotic checkpoint controls and
153 missegregation[24,25]. Taken together, these data provide independent evidence for BFB arising as a
154 result of oncogenic RAS signaling and telomere shortening as the underlying mechanism for CN
155 signature 1.

156 We applied these approaches to the remaining signatures to identify statistically significant
157 genomic associations using a false discovery rate <0.05 (Figure 2b-e, Figure 3, Supplementary
158 Figures 5, 6, 7, 8 and 9, Supplementary Tables 3, 4, 5, 6, 7 and 8).

159 CN signature 2 showed frequent breakpoints per 10MB, single changes in copy-number (resulting
160 in 3 copies), chains of oscillating copy-number, and was significantly correlated with tandem
161 duplicator phenotype scores (r=0.3, P=0.004) and SNV signature 5 (r=0.26, P=0.02). In addition,
162 this signature was enriched in patients with mutations in *CDK12* (P=0.02, Mann-Whitney test,
163 Supplementary Table 6), in keeping with previous studies that have demonstrated large tandem
164 duplication in cases with inactivating *CDK12* mutations[26].

165 CN signature 4 was characterised by high copy-number states (4-8 copies) and predominant copy-
166 number change-points of size 2. This pattern indicates a mutational process of late whole-genome
167 duplication (WGD)[27]. Significantly increased signature 4 exposure in cases with aberrant PI3K/AKT
168 signaling provided further support for late WGD as oncogenic PIK3CA induces tolerance to
169 genome doubling[28] (P=2e-22, Mann-Whitney test, mutation of *PIK3CA* or amplification of *AKT,*
170 *EGFR, MET, FGFR3* and *ERBB2).* Signature 4 was also seen at higher levels in cases with
171 mutations in genes encoding proteins from Toll-like receptor signaling cascades (P=2e-07),
172 interleukin signaling pathways (P=3e-24) and *CDK12* (P=0.0009), as well as those with amplified
173 *CCNE1* (P=2e-10) and *MYC* (P=9e-12). It was also significantly correlated with telomere length
174 (r=0.46, P=4e-05).

175 CN signature 6 showed extremely high copy-number states and high copy-number change-points
176 for small segments interspersed among larger, lower-copy segments. This suggests a mutational
177 process resulting in focal amplification. Increased signature 6 exposure was associated with
178 mutations in genes encoding proteins across diverse pathways, including aberrant G1/S cell cycle
179 checkpoint control (through either amplification of *CCNE1*, *CCND1*, *CDK2, CDK4* or *MYC*,
180 deletion/inactivation of *RB1* or mutation in *CDK12*), Toll-like receptor signaling cascades and
181 PI3K/AKT signaling (P<0.05). However, as many of these statistical associations are marked by
182 gene amplification, it is difficult to determine whether the copy number states represent causal
183 events or are simply a consequence of focal amplification. Exposure to CN signature 6 was also
184 positively correlated with age at diagnosis (r=0.31, P=6e-12) and age-related SNV signature 1[14]
185 (r=0.43, P=3e-06).

186 CN signature 5 was significantly associated with predicted chromothriptic-like events using the
187 Shatterproof algorithm[29] (r=0.44, P=2e-03). Chromothripsis is considered rare in HGSOC[12,27,30].
188 However, the key component of this signature—the presence of copy-number change points
189 centered at 0.5 copies—suggests that the events are subclonal. This implies that chromothripsis
190 may be an underestimated oncogenic mechanism in HGSOC that could reflect ongoing formation
191 and rupture of micronuclei[31].

192 CN signature 3 was characterized by an even distribution of breaks across all chromosomes, and
193 copy number changes from diploid to single copy (LOH). CN signature 3 was significantly enriched
194 in cases with mutations in *BRCA1* and *BRCA2,* and other HR genes including *BARD1, PALB2* and
195 *ATR* (P=0.002, Mann-Whitney test). It was also correlated with the HRD-related SNV signature 3
196 (r=0.32, P=0.002) and anti-correlated with age at diagnosis and age-related SNV signature 1
197 (P<0.05). CN signature 3 was also enriched in cases with loss of function mutations in *PTEN*
198 (P=0.002, Mann-Whitney test). Taken together, these data suggest that CN signature 3 is driven
199 by BRCA1/2-related HRD mechanisms.

200 CN signature 7, like CN signature 3, also demonstrated an even distribution of breaks across all
201 chromosomes. By contrast with CN signature 3, single copy-number changes were observed from

202 a tetraploid rather than a diploid state (Figure 3). Although there was correlation with the HRD-
203 related SNV signature 3, there was no enrichment with *BRCA1/2* mutation, suggesting alternative
204 HRD mechanisms as potential mutational processes.

205 We also investigated relationships between CN signatures. BRCA1 dysfunction and *CCNE1*
206 amplification have been shown to be mutually exclusive in HGSOC[32], and we observed that CN
207 signature 3 (*BRCA1/2* HRD) and CN signature 6 (marked by aberrant G1/S cell cycle checkpoint
208 control) showed mutually exclusive associations (Figure 2b-e). Loss of *BRCA1* and *BRCA2* are
209 early driver events in HGSOC, and to investigate acquisition of additional mutational processes,
210 we studied four BriTROC-1 cases with deleterious germline *BRCA2* mutations and confirmed
211 somatic loss of heterozygosity at *BRCA2* (Figure 4). A diverse and variable number of CN
212 signatures was seen in these cases, including substantial exposures to CN signature 1 (RAS
213 signaling) in three of the four cases.

214 Copy-number signatures predict overall survival

215 We next explored the association between individual CN signature exposures and overall survival
216 using a combined dataset of 575 diagnostic samples with clinical outcomes. We trained a
217 multivariate Cox proportional hazards model on 417 cases and tested this on the remaining 158
218 cases (Figure 5, Supplementary Table 9). CN signature exposure was significantly predictive of
219 survival (Training: P=0.002, log-rank test; stratified by age and cohort; Test: P=0.05, C-index=0.56,
220 95% CI:0.50-0.62; Entire cohort: P=0.002, log-rank test; stratified by age and cohort). Across the
221 entire cohort, poor outcome was significantly predicted by CN signature 1 (P=0.0008) and CN
222 signature 2 exposures (P=0.03), whilst good outcome was significantly predicted by exposures to
223 CN signatures 3 (P=0.05) and 7 (P=0.006).

224 Unsupervised hierarchical clustering of samples by signature exposures identified three clusters
225 (Figure 5). Despite showing significant survival differences (P=0.004, log-rank test; stratified by
226 age and cohort), these clusters did not provide any prognostic information in addition to that
227 identified from the Cox proportional hazards model; cluster 2 was dominated by patients with high
228 signature 1 exposures (poor prognosis), cluster 3 showed high signature 3 exposures (good
229 prognosis) and cluster 1 had mixed signature exposures (Supplementary Figure 10).

230 Copy-number signatures indicate relapse following chemotherapy

231 Using a generalised linear model, we investigated whether copy-number signatures could be used
232 to predict outcome following chemotherapy across 36 patients from the BriTROC-1 study with
233 paired diagnostic and relapse samples[16]. The model showed CN signature 1 exposures at the time
234 of diagnosis to be significantly predictive of platinum-resistant relapse (P=0.02, z-test,
235 Supplementary Table 10).

236 Using the same 36 sample pairs, we also investigated whether chemotherapy treatment changed
237 CN signature exposures. No significant effects on exposures were observed following
238 chemotherapy treatment using a linear model that accounted for signature exposure at time of
239 diagnosis, number of lines of chemotherapy and patient age (P>0.05, F-test, Supplementary Table
240 10). The only variable showing a significant association with exposure at relapse was signature
241 exposure at diagnosis (P<0.01, F-test, Supplementary Table 11).

242 # Discussion

243 Copy-number signatures provide a framework that is able to rederive the major defining elements
244 of HGSOC genomes, including defective HR[8], amplification of *CCNE1*[9] and amplification-
245 associated fold-back inversions[13]. In addition, the CN signatures show significant associations with
246 known driver gene mutations in HGSOC and provide the ability to detect novel associations with
247 gene mutations. We derived signatures using inexpensive shallow whole genome sequencing of
248 DNA from core biopsies. These approaches are rapid and cost effective, thus providing a clear
249 path to clinical implementation. Copy-number signatures open new avenues for clinical trial design
250 by highlighting contributions from underlying mutational processes that depend on oncogenic RAS
251 and PI3K/AKT signaling.

252 We found that almost all patients with HGSOC demonstrated a mixture of signatures indicative of
253 combinations of mutational processes. These results suggest that early *TP53* mutation, the
254 ubiquitous initiating event in HGSOC, may permit multiple mutational processes to co-evolve,
255 potentially simultaneously. Although further work is needed to define the precise timing of
256 signature exposures, early driver events such as *BRCA2* mutation still permit a diverse and
257 variable number of CN signatures in addition to an HRD signature (Figure 4). These additional
258 signature exposures may alter the risk of developing therapeutic resistance, particularly when only
259 a single mutational process such as HRD is targeted.

260 High exposure to CN signature 3, characterised by BRCA1/2-related HRD, is associated with
261 improved overall survival, confirming prior data showing that *BRCA1/2* mutation is associated with
262 long survival in HGSOC[33,34]. Conversely, high exposure to signature 1, which is characterised by
263 oncogenic RAS signaling (including *NF1, KRAS* and *NRAS* mutation), predicts subsequent
264 platinum-resistant relapse and poor survival. This suggests that powerful intrinsic resistance
265 mechanisms are present at the time of diagnosis and can be readily identified using CN signature
266 analysis. This hypothesis is supported by the presence of exposure to CN signature 1 in germline
267 *BRCA2*-mutated cases (Figure 4) as well as our previous work demonstrating the expansion of a
268 resistant subclonal *NF1*-deleted population following chemotherapy treatment in HGSOC[35] and
269 poor outcomes in *Nf1*-deleted murine models of HGSOC[36]. Our CN signature analysis of *BRCA2*-
270 mutated cases also concurs with PCAWG/ICGC data showing that over half (9/16) of *NF1*-mutated
271 cases also harboured mutations in *BRCA1* or *BRCA2*[12]. These data suggest a complex interplay
272 between RAS signaling and HRD. Thus, RAS signaling may be an important target, especially in
273 first line treatment, to prevent emergence of platinum-resistant disease.

274 We found that CN signature exposures were not significantly altered between diagnosis and
275 disease relapse in 36 sample pairs with a median interval of 30.6 months[16]. This suggests that the
276 underlying mutational processes in HGSOC are relatively stable and that genome-wide patterns of
277 copy-number change mainly reflect historic alterations to the genome acquired during

278 tumorigenesis[37]. Relative invariant genomic changes were also observed in the ARIEL2 trial,
279 where genome-wide loss-of-heterozygosity was used to predict HRD, and only 14.5% (17/117)
280 cases changed LOH status between diagnosis and relapse[7].

281 Larger association studies will be required to further refine CN signature definitions and
282 interpretation. The application of our approach to other tumour types is likely to extend the set of
283 signatures beyond the robust core set identified here. Basal-like breast cancers, squamous cell
284 and small cell lung carcinoma, which all have high rates of *TP53* mutation and genomic instability[2],
285 are promising next targets. Although it is likely that the strong associations have identified the
286 driver mutational processes for CN signatures 1 and 3, functional studies will be required to
287 establish causal links for the remaining signatures. For example, CN signature 6 was significantly
288 associated with multiple mutated pathways, and this association was primarily driven by
289 amplification of target genes. As this signature represented focal amplification events, it is difficult
290 to determine whether amplification of specific genes drives the underlying mutational process or
291 the amplifications emerge as a consequence of strong selection of advantageous phenotypes. Our
292 data does not provide timing information for exposures and there is the real possibility that one
293 mutational process may well drive the emergence of other mutational processes. For example, the
294 association between signature 6 and PI3K signalling is also shared with signature 4.

295 Other limitations of this work are technical: we integrated data from three sources, using three
296 different pre-processing pipelines, and the ploidy determined by different pipelines can have a
297 significant effect on the derived signatures. For example, high-ploidy CN signature 4 was
298 predominantly found in the sequenced samples that underwent careful manual curation to identify
299 whole-genome duplication events. When extending to larger sample sets, a unified processing
300 strategy with correct ploidy determination is likely to produce improved signature definitions.
301 Another technical limitation is the resolution of copy-number calling from sWGS (limited to 30kb
302 bins) and future application to large cohorts of deeply sequenced samples will be needed to
303 improve the resolution of the CN signatures.

304

305 Efforts to identify discrete, clinically relevant subtypes of disease have been successful in many
306 cancer types[38-40]. However, HGSOC lacks clinically-relevant patient stratification, which is reflected
307 in continued poor survival. We show that HGSOC genomes are shaped by multiple mutational
308 processes that preclude simple subtyping. Thus, our results suggest that HGSOC is a continuum
309 of genomes. By dissecting the mutational forces shaping HGSOC genomes, our study paves the
310 way to understanding extreme genomic complexity, as well as revealing the evolution of tumors as
311 they relapse and acquire resistance to chemotherapy.

Accession Codes

EGAS00001002557

Acknowledgements

The BriTROC-1 study was funded by Ovarian Cancer Action (to IMcN and JDB, grant number 006). We would like to acknowledge funding and support from Cancer Research UK (grant numbers A15973, A15601, A18072, A17197, A19274 and A19694), the Universities of Cambridge and Glasgow, National Institute for Health Research Cambridge and Imperial Biomedical Research Centres, National Cancer Research Network, the Experimental Cancer Medicine Centres at participating sites, the Beatson Endowment Fund and Hutchison Whampoa Limited. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the Biorepository, Bioinformatics, Histopathology and Genomics Core Facilities of the Cancer Research UK Cambridge Institute and the Pathology Core at the Cancer Research UK Beatson Institute for technical support. We would like to thank members of PCAWG Evolution and Heterogeneity Working Group for the consensus copy-number analysis, PCAWG Structural Variation Working Group for the consensus structural variants and PCAWG Technical Working Group for annotating driver mutations in the 112 PCAWG-OV samples.

Author contributions

G.M., T.E.G., F.M., I.McN., J.D.B. conceptualized the study; S.D., R.M.G., M.L., E.B., A.M., A.W., S.S., R.E., G.D.H., A.C., C.G., M.H., C.F., H.G., D.M., A.Ho., G.B., I.McN., J.D.B. conducted sample collection; T.E.G., D.E., A.M.P., L.A.L., A.Ha., C.W., C.N., L.Mi., L.N.S., M.J.L., L.Mo., A.S., J.P. performed experiments; G.M., T.E.G., D.D.S., M.E., D.S., B.Y., O.H., F.M. performed data analysis; G.M., D.D.S., F.M. developed the methodology and software; G.M., T.E.G., D.D.S., F.M., I.McN., J.D.B. wrote the manuscript.

Competing Financial Interests Statement

The following authors the authors have a competing interest as defined by Nature Research:
C.G. Personal interest: Roche, AstraZeneca, Tesaro, Clovis, Foundation One, Nucana. Research funding: AstraZeneca, Novartis, Aprea, Nucana, Tesaro. Named co-inventor on five patents (issued: PCT/US2012/040805; pending: PCT/GB2013/053202, 1409479.1, 1409476.7 and 1409478.3)
H.G. Employment: AstraZeneca
I.McN. Personal interest: Clovis Oncology.
J.D.B. Cofounder and shareholder of Inivata Ltd (a cancer genomics company that commercializes ctDNA analysis)
All other authors declare that they have no competing financial or non-financial interests as defined by Nature Research.

350 # References

351 1. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat*
352 *Genet* **45**, 1127-33 (2013).
353 2. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular
354 classification within and across tissues of origin. *Cell* **158**, 929-44 (2014).
355 3. Ahmed, A.A. *et al.* Driver mutations in TP53 are ubiquitous in high grade serous carcinoma
356 of the ovary. *J Pathol* **221**, 49-56 (2010).
357 4. Vaughan, S. *et al.* Rethinking ovarian cancer: recommendations for improving outcomes.
358 *Nat. Rev. Cancer* **11**, 719-725 (2011).
359 5. Fong, P.C. *et al.* Poly(ADP)-Ribose Polymerase Inhibition: Frequent Durable Responses in
360 BRCA Carrier Ovarian Cancer Correlating With Platinum-Free Interval. *J. Clin. Oncol.* **28**,
361 2512-2519 (2010).
362 6. Gelmon, K.A. *et al.* Olaparib in patients with recurrent high-grade serous or poorly
363 differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre,
364 open-label, non-randomised study. *Lancet Oncol.* **12**, 852-861 (2011).
365 7. Swisher, E.M. *et al.* Rucaparib in relapsed, platinum-sensitive high-grade ovarian
366 carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. *Lancet*
367 *Oncol* **18**, 75-87 (2017).
368 8. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
369 9. Etemadmoghadam, D. *et al.* Integrated genome-wide DNA copy number and expression
370 analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas.
371 *Clin. Cancer Res.* **15**, 1417-1427 (2009).
372 10. Verhaak, R.G. *et al.* Prognostically relevant gene signatures of high-grade serous ovarian
373 carcinoma. *J Clin Invest* **123**, 517-25 (2013).
374 11. Chen, G.M. *et al.* Consensus on Molecular Subtypes of Ovarian Cancer. *bioRxiv* (2017).
375 12. Patch, A.-M. *et al.* Whole–genome characterization of chemoresistant ovarian cancer.
376 *Nature* **521**, 489-494 (2015).
377 13. Wang, Y.K. *et al.* Genomic consequences of aberrant DNA repair mechanisms stratify
378 ovarian cancer histotypes. *Nat Genet* **49**, 856-865 (2017).
379 14. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
380 415-21 (2013).
381 15. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome
382 sequences. *Nature* **534**, 47-54 (2016).
383 16. Goranova, T. *et al.* Safety and utility of image-guided research biopsies in relapsed high-
384 grade serous ovarian carcinoma-experience of the BriTROC consortium. *Br J Cancer* **116**,
385 1294-1301 (2017).
386 17. Campbell, P.J. *et al.* Pan-cancer analysis of whole genomes. in *bioRxiv* (2017).
387 18. Murnane, J.P. Telomere dysfunction and chromosome instability. *Mutat Res* **730**, 28-36
388 (2012).
389 19. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes.
390 *Cell* **152**, 1226-36 (2013).
391 20. Ng, C.K. *et al.* The role of tandem duplicator phenotype in tumour evolution in high-grade
392 serous ovarian cancer. *J Pathol* **226**, 703-12 (2012).
393 21. Menghi, F. *et al.* The tandem duplicator phenotype as a distinct genomic configuration in
394 cancer. *Proc Natl Acad Sci U S A* **113**, E2373-82 (2016).
395 22. Lee, M. *et al.* Comparative analysis of whole genome sequencing-based telomere length
396 measurement techniques. *Methods* **114**, 4-15 (2017).
397 23. Zakov, S., Kinsella, M. & Bafna, V. An algorithmic approach for breakage-fusion-bridge
398 detection in tumor genomes. *Proc Natl Acad Sci U S A* **110**, 5546-51 (2013).
399 24. Knauf, J.A. *et al.* Oncogenic RAS induces accelerated transition through G2/M and
400 promotes defects in the G2 DNA damage and mitotic spindle checkpoints. *J Biol Chem*
401 **281**, 3800-9 (2006).
402 25. Saavedra, H.I., Fukasawa, K., Conn, C.W. & Stambrook, P.J. MAPK mediates RAS-
403 induced chromosome instability. *J Biol Chem* **274**, 38083-90 (1999).

404 26. Popova, T. *et al.* Ovarian Cancers Harboring Inactivating Mutations in CDK12 Display a
405 Distinct Genomic Instability Pattern Characterized by Large Tandem Duplications. *Cancer*
406 *Res* **76**, 1882-91 (2016).
407 27. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**,
408 1134-40 (2013).
409 28. Berenjeno, I.M. *et al.* Oncogenic PIK3CA induces centrosome amplification and tolerance
410 to genome doubling. *Nat Commun* **8**, 1773 (2017).
411 29. Govind, S.K. *et al.* ShatterProof: operational detection and quantification of chromothripsis.
412 *BMC Bioinformatics* **15**, 78 (2014).
413 30. Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex
414 rearrangements spawned by homology-independent mechanisms. *Genome Res* **23**, 762-
415 76 (2013).
416 31. Bakhoum, S.F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA
417 response. *Nature* **553**, 467-472 (2018).
418 32. Etemadmoghadam, D. *et al.* Synthetic lethality between CCNE1 amplification and loss of
419 BRCA1. *Proc Natl Acad Sci U S A* **110**, 19489-94 (2013).
420 33. Candido Dos Reis, F.J. *et al.* Germline mutation in BRCA1 or BRCA2 and ten-year survival
421 for women diagnosed with epithelial ovarian cancer. *Clin Cancer Res* **21**, 652-7 (2015).
422 34. Norquist, B.M. *et al.* Mutations in Homologous Recombination Genes and Outcomes in
423 Ovarian Carcinoma Patients in GOG 218: An NRG Oncology/Gynecologic Oncology Group
424 Study. *Clin Cancer Res* **24**, 777-783 (2018).
425 35. Schwarz, R.F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian
426 cancer: a phylogenetic analysis. *PLoS Med* **12**, e1001789 (2015).
427 36. Walton, J.B. *et al.* CRISPR/Cas9-derived models of ovarian high grade serous carcinoma
428 targeting Brca1, Pten and Nf1, and correlation with platinum sensitivity. *Scientific Reports*
429 **7**, 16827 (2017).
430 37. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv* (2017).
431 38. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours
432 reveals novel subgroups. *Nature* **486**, 346-52 (2012).
433 39. Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature*
434 **497**, 67-73 (2013).
435 40. Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically
436 distinct subgroups with therapeutic relevance. *Nat Genet* **48**, 1131-41 (2016).

437

438 # Figure Legends

439 **Figure 1 | Copy-number signature identification from shallow whole genome sequence data**
440 **and validation in independent cohorts**

441 **a.** Step 1: Absolute copy-numbers are derived from sWGS data; Step 2: genome-wide distributions
442 of six fundamental copy-number features are computed; Step 3: Gaussian or Poisson mixture
443 models (depending on data type) are fitted to each distribution and the optimal number of
444 components is determined (ranging from 3–10) ; Step 4: the data are represented as a matrix with
445 36 mixture component counts per tumor. Step 5: Non-negative matrix factorization is applied to the
446 components-by-tumor matrix to derive the tumor-by-signature matrix and the signature-by-
447 components matrix.

448 **b.** Heat maps show component weights for copy number signatures in two independent cohorts of
449 HGSOC samples profiled using WGS and SNP array. Correlation coefficients are provided in
450 Supplementary Table 2.

451 **Figure 2 | Linking copy-number signatures with mutational processes**

452 **a** Component weights for copy number signature 1. Barplots (upper panel) are grouped by copy
453 number feature and show weights for each of the 36 components. The middle panel shows the
454 mixture model distributions which are shaded by the component weight - solid colours have a high
455 weight and transparent have low weight (contrasting colours are randomly assigned). Lower panel
456 shows genome-wide distribution (histogram or density) of each copy number feature, across the
457 BriTROC-1 cohort, with coloured plots indicating important distributions (> 0.1 component weight).
458 (Note: similar plots for other CN signatures are shown in Figure 3 and Supplementary Figure 5).

459 **b** Associations between CN signature exposures and other features. Purple indicates positive
460 correlation and orange negative correlation (see also Supplementary Figure 6). Numbers at the
461 right of the panel indicate cases included in each analysis. Only significant correlations are shown
462 ($P<0.05$).

463 **c** Associations between CN signature exposures and SNV signatures. Purple indicates positive
464 correlation and orange negative correlation (see also Supplementary Figure 6). The number at the
465 right of the panel indicates cases included in the analysis.

466 **d and e** Difference in CN signature exposures between cases with mutations in specific genes (**d**)
467 and mutated/wildtype reactome pathways (**e**). The absolute difference in mean signature
468 exposures was calculated for cases with and without mutations. Colors in filled circles indicate
469 extent of difference. Only differences with FDR $P<0.05$ (Mann-Whitney test) are shown (see also
470 Supplementary Figure 7).

471 Numbers at the right of the panel indicate cases with mutations (SNVs, amplifications or deletions)
472 in each gene/pathway.

473 **Figure 3 | The seven copy-number signatures in HGSOC**

474 Description of the defining component weights, key associations and proposed mechanisms for the
475 seven copy number signatures.
476 *only the top three mutated genes for each of the pathways associated with CN signatures 4, 6
477 and 7 are shown (the list of all significant genes is provided in Supplementary Tables 7 and 8).

478 **Figure 4 | CN signature exposures of four BriTROC-1 patients with germline *BRCA2***
479 **mutations and somatic loss of heterozygosity**

480 Stacked bar plots show copy-number signature exposures for four BriTROC-1 cases with
481 pathogenic germline *BRCA2* mutations and confirmed somatic loss of heterozygosity (LOH) at the
482 *BRCA2* locus.

483 **Figure 5 | Association of survival with copy-number signatures**

484 Upper panel: Stacked barplots show CN signature exposures for each patient. Patients were
485 ranked by risk of death estimated by a multivariate Cox proportional hazards model stratified by
486 age and cohort, with CN signature exposures as covariates.
487 Middle panel: The matrix indicates group for each patient assigned by unsupervised clustering of
488 CN signature 1, 2, 3 and 7 exposures (see also Supplementary Figure 10).
489 Lower panel: Linear fit of signature exposures ordered by risk predicted by the Cox proportional
490 hazards model.
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513

514 # Online Methods

515 ## Patients and samples

516 The BriTROC-1 study has been described previously[16]. Characteristics of the 142 patients
517 included in this study are given in Supplementary Table 1. The study is sponsored by NHS Greater
518 Glasgow and Clyde and ethics/IRB approval was given by Cambridge Central Research Ethics
519 Committee (Reference 12/EE/0349). The study enrolled patients with recurrent ovarian high-grade
520 serous or grade 3 endometrioid carcinoma who had relapsed following at least one line of
521 platinum-based chemotherapy and whose disease was amenable either to image-guided biopsy or
522 secondary debulking surgery. At study entry, patients were classified as having either platinum-
523 sensitive relapse (i.e. relapse six months or more following last platinum chemotherapy) or
524 platinum-resistant relapse (i.e. relapse less than six months following prior platinum chemotherapy)
525 (Supplementary Figure 2). All patients provided written informed consent. Access to archival
526 diagnostic formalin-fixed tumor was also required. Survival was calculated from the date of
527 enrolment to the date of death or the last clinical assessment, with data cutoff at 1 December
528 2016. At subsequent relapse or progression after chemotherapy following study entry, patients
529 could optionally have a second biopsy under separate consent.
530 DNA was extracted from 300 samples of 142 patients - 158 methanol-fixed relapse biopsies and
531 142 FFPE archival diagnostic tissues. Germline DNA was extracted from blood samples of 137
532 patients.

533 ## Tagged-amplicon sequencing

534 Mutation screening of *TP53, PTEN, EGFR, PIK3CA, KRAS* and *BRAF* was performed on all 300
535 samples using tagged-amplicon sequencing as previously described[16]. DNA extracted from blood
536 was analyzed by tagged-amplicon sequencing for *BRCA1* and *BRCA2* germline mutations.

537 ## Shallow whole genome sequencing (sWGS)

538 Libraries for sWGS were prepared from 100ng DNA using modified TruSeq Nano DNA LT Sample
539 Prep Kit (Illumina) protocol[41]. Quality and quantity of the libraries were assessed with DNA-7500 kit
540 on 2100 Bioanalyzer (Agilent Technologies) and with Kapa Library Quantification kit (Kapa
541 Biosystems) according to the manufacturer's protocols. Sixteen to twenty barcoded libraries were
542 pooled together in equimolar amounts and each pool was sequenced on HiSeq4000 in SE-50bp
543 mode.
544 Prior to sequencing we estimated the required sequencing depth by adapting calculations made in
545 previous work that explored the relationship between sequencing depth (reads per sample) and
546 copy number calling accuracy[42]. Based on these analyses, we devised a power calculator for
547 sWGS copy number analysis (see URL 1, described in [43]). We estimated that with an average

548 ploidy of 3 and purity of 0.65, a sequencing depth of at least 2.7 million reads is required to detect
549 single, clonal copy-number changes (minimum 60kb) at 90% power and alpha 0.05. After analysis
550 we determined that BritROC 3-star samples had an average purity of 0.66, ploidy of 2.7, and were
551 sequenced to an average depth of 8.6 million reads. This allowed us to detect single copy-number
552 changes with 90% power, and alpha 0.05 down to subclonal frequencies of 55%.

553 Deep whole genome sequencing

554 Deep whole-genome sequencing was performed on 56 tumors with confirmed *TP53* mutations and
555 matched normal samples, of which 48 passed quality control. Libraries were constructed with
556 ~350-bp insert length using the TruSeq Nano DNA Library prep kit (Illumina) and sequenced on an
557 Illumina HiSeq X Ten System in paired-end 150-bp reads mode. The average depth was 60×
558 (range 40-101×) in tumors and 40× (range 24-73×) in matched blood samples.

559 Variant calling

560 Read alignment and variant calling of tagged-amplicon sequencing data were processed as
561 described[41]. Deep WGS samples were processed with bcbio-nextgen[44] using Ensemble somatic
562 variants called by two methods out of VarDict[45], Varscan[46] and FreeBayes[47]. Somatic SNV calls
563 were further filtered based on mapping quality, base quality, position in read, and strand bias as
564 described[40]. In addition, the blacklisted SNVs from the Sanger Cancer Genomics Project pipeline
565 derived from a panel of unmatched normal samples were used for filtering[48].

566 Data download

567 PCAWG-OV: Consensus SNVs and INDELs (October 2016 release), consensus structural variants
568 (v 1.6), consensus copy-number calls (January 2017 release), donor clinical (August 2016 v7-2)
569 and donor histology information (August 2016 v7) for 112 ovarian cancer samples were
570 downloaded from the PCAWG data portal. ABSOLUTE[49] copy-number calls were used for
571 analysis.
572 TCGA: ABSOLUTE[49] copy-number profiles from Zack et al[27] for 415 ovarian cancer TCGA
573 samples were downloaded from Synapse[50]. SNVs for these samples were downloaded from the
574 Broad Institute TCGA Genome Data Analysis Center (Broad Institute TCGA Genome Data
575 Analysis Center: Firehose stddata__2016_01_28 run. doi:10.7908/C11G0KM9, Broad Institute of
576 MIT and Harvard). Donor clinical data were downloaded from the TCGA data portal.

577 Absolute copy-number calling from sWGS

578 *Segmentation:* sWGS reads were aligned and relative copy-number called as described[41]. After
579 inspection of the *TP53* mutation status and relative copy-number profiles of the 300 sequenced
580 BriTROC-1 samples, 47 were excluded from downstream analysis for the following reasons: low

581 purity (24), mislabeled (7), pathology re-review revealed sample was not HGSOC (3), no
582 detectable *TP53* mutation (13). Of the 253 BriTROC-1 samples analysed, 111 were FFPE-fixed.
583 Fifty seven out of 253 showed an over segmentation artefact (likely due to fixation). A more strict
584 segmentation was subsequently applied to these samples to yield a usable copy-number profile.
585

586 *Absolute copy number:* We combined relative copy-number profiles generated by QDNAseq[42] with
587 mutant allele frequency identified using tagged amplicon sequencing in a probabilistic graphical
588 modelling approach to infer absolute copy-number profiles. Using Expectation-Maximisation, the
589 model generated a posterior over a range of *TP53* copy-number states, using the *TP53* mutant
590 allele frequency to estimate purity for each state. The *TP53* copy-number state that provided the
591 highest likelihood of generating a clonal absolute copy-number profile was used to determine the
592 final absolute copy-number profile. To test the validity of this approach, we compared purity and
593 ploidy estimates derived from sWGS to those derived from 60× WGS using the Battenberg
594 algorithm for copy-number calling[51]. Pearson correlation coefficients were computed for both ploidy
595 and purity estimates using 34 3-star (see *Quality rating*) BriTROC-1 samples with matched sWGS
596 and WGS (Supplementary Figure 11).
597

598 *Quality rating:* Following absolute copy-number fitting, samples were rated using a 1-3 star system.
599 1-star samples (n=54) showed a noisy copy-number profile and were considered likely to have
600 incorrect segments and missing calls. These were excluded from further analysis. 2-star samples
601 (n=52) showed a reasonable copy-number profile with only a small number of miscalled segments.
602 These samples were used (with caution) for some subsequent analyses. 3-star samples (n=147)
603 showed a high-quality copy-number profile that was used in all downstream analyses. The
604 maximum star rating observed per patient was 1-star in 15 patients, 2-star in 26, and 3-star in 91
605 patients. Seventy-two out of 111 FFPE-fixed samples (64%) were amenable to signature analysis.
606 This is consistent with typical sequencing success rates for archival material[52].

607 Copy-number signature identification

608 *Preprocessing:* 91 3-star BriTROC-1 absolute copy-number profiles were summarized using the
609 genome-wide distribution of six different features (outlined in Figure 1):
610 1. Segment size - the length of each genome segment;
611 2. Breakpoint count per 10MB - the number of genome breaks appearing in 10MB sliding
612 windows across the genome;
613 3. Change-point copy-number - the absolute difference in CN between adjacent segments
614 across the genome;
615 4. Segment copy-number - the observed absolute copy-number state of each segment;
616 5. Breakpoint count per chromosome arm - the number of breaks occurring per chromosome
617 arm;

618 6. Length of segments with oscillating copy-number - a traversal of the genome counting the
619 number of contiguous CN segments alternating between two copy-number states, rounded to
620 the nearest integer copy-number state.
621
622 *Mixture modelling:* For each of the feature density distributions, we applied mixture modelling to
623 identify its distinct components. For distributions representing segment-size, change-point copy-
624 number, and segment copy-number we employed mixtures of Gaussians. For distributions
625 representing breakpoint count per 10MB, length of segments with oscillating copy-number, and
626 breakpoint count per chromosome arm we employed mixtures of Poissons. Mixture modelling was
627 performed using the FlexMix V2 package in R[53]. The algorithm was run for each distribution with
628 the number of components ranging from 2-10. The optimal number of components was selected as
629 the run showing the lowest Bayesian Information Criterion, resulting in a total of 36 components
630 (see Figure 1 and Supplementary Table 3 for breakdown). Next, for each copy-number event, we
631 computed the posterior probability of belonging to a component. For each sample, these posterior
632 event vectors were summed resulting in a sum-of-posterior probabilities vector. All sum-of-
633 posterior vectors were combined in a patient-by-component sum-of-posterior probabilities matrix.
634
635 *Signature identification:* The NMF Package in R[54], with the Brunet algorithm specification[55] was
636 used to deconvolute the patient-by-component sum-of-posteriors matrix into a patient-by-signature
637 matrix and a signature-by-component matrix. A signature search interval of 3-12 was used, running
638 the NMF 1000 times with different random seeds for each signature number. As provided by the
639 NMF Package[54], the cophenetic, dispersion, silhouette, and sparseness coefficients were
640 computed for the signature-by-component matrix (basis), patient-by-signature matrix (coefficients)
641 and connectivity matrix (consensus, representing patients clustered by their dominant signature
642 across the 1000 runs). 1000 random shuffles of the input matrix were performed to get a null
643 estimate of each of the scores (Supplementary Figure 3). We sought the minimum signature
644 number that yielded stability in the cophenetic, dispersion and silhouette coefficients, and that
645 yielded the maximum sparsity which could be achieved without exceeding that which was
646 observed in the randomly permuted matrices. As a result, 7 signatures were deemed optimal under
647 these constraints and were chosen for the remaining analysis.
648
649 *Signature assignment:* For the remaining 26 2-star patient samples, and the 82 secondary patient
650 samples (from patients with 2- or 3-star profiles from additional tumor samples), the LCD function
651 in the YAPSA package in Bioconductor[56] was used to assign signature exposures.

652 Copy-number signature validation

653 The signature identification procedure described above was applied to copy-number profiles from
654 two independent datasets: 112 whole-genome sequenced (approximately 40×) HGSOC samples

655 processed as part of ICGC Pan-Cancer Analysis of Whole Genomes Project[17], (denoted here as
656 PCAWG-OV) and 415 SNParray profiling of HGSOC cases as part of TCGA[27]. The number of
657 signatures was fixed at 7 for matrix decomposition with NMF. Pearson correlation was computed
658 between the BriTROC-1 signature-by-component weight matrix and each of the PCAWG-OV and
659 TCGA signature-by-component matrices, signature by signature (Supplementary Table 2).

660 Association of copy-number signature exposures with other features

661 Association of signature exposures with other features was performed using one of two
662 procedures: for a continuous association variable, correlation was performed; for a binary
663 association variable, patients were divided into two groups and a Mann-Whitney test was
664 performed to test for differences in signature exposure medians between the two groups. A more
665 detailed explanation of each of these association calculations is given below. (Note: of the 48 deep
666 WGS BriTROC-1 samples that passed QC, only 44 had matched 2- and 3-star sWGS copy-
667 number profiles. As signature exposures from sWGS were used for BriTROC-1 sample
668 associations, only these 44 samples could be used).
669

670 *Age at diagnosis.* Patient age at diagnosis for 112 PCAWG-OV samples and 415 TCGA samples
671 was used to compute Pearson correlation with signature exposures.
672

673 *Amplification associated fold-back inversions.* For 111 PCAWG-OV samples, the fraction of
674 amplification associated fold-back inversion events per sample was calculated as the proportion of
675 head-to-head inversions (h2hINVs) within a 100kb window amplified region (copy number ≥5)
676 relative to the total number of SV calls per sample. 94 samples had at least 1 h2hINV event out of
677 which 58 had h2hINV events in amplified regions. On average they accounted for 4% of SV calls.
678 As these are rare events, only samples showing a non-zero fraction of fold-back inversions (n=67)
679 were used to compute Pearson correlation with signature exposures.
680

681 *Telomere length.* Telomere lengths of 44 deep WGS tumor samples from the BriTROC-1 cohort
682 were estimated using the Telomerecat algorithm[57]. Telomere length estimates ranged from 1.5kb -
683 11kb with an average of 4kb. Correlation between telomere length and copy-number signature
684 exposures was calculated with age and tumor purity as covariates using the ppcor package in R[58].
685

686 *Chromothripsis.* Copy-number and translocation information from 111 PCAWG-OV samples were
687 used to detect chromothripsis-like events using the Shatterproof software with default

688 parameters[29]. Shatterproof, a state-of-the-art software, incorporates a wide range of hallmarks of
689 chromothripsis in its detection algorithm as a precise definition of chromothripsis remains elusive.
690 Govind et al. recommend a threshold of 0.37 based on their observations that normal samples
691 produced a low number of calls with low scores (maximum 0.37) while prostate, colorectal and
692 small cell lung cancer samples that were known to have chromothriptic events, produced the
693 highest scores [29]. Previous studies have reported a low incidence of chromothriptic events in
694 HGSOC [12,27,30]. The number of calls per sample in the PCAWG-OV samples ranged from 5 to 47
695 with an average of 23. The score per call ranged from 0.15-0.62 with a median of 0.38. Therefore,
696 a conservative threshold was set at the 95th percentile of our distribution of scores to minimise
697 false positives and calls with scores greater than 0.48 were used to obtain a count of
698 chromothriptic events per sample. As chromothriptic events are rare in HGSOC, only samples
699 showing a non-zero number of events (n=61) were used to compute Pearson correlation with
700 signature exposures. Of 61 samples with scores above the threshold, 49 (80.3%) had 1-2 events,
701 11 samples (18%) had 3-6 events and 1 sample (1.6%) had 10 events.
702
703 *Tandem duplicator phenotypes.* Tandem duplicator phenotype (TDP) scores were calculated for
704 111 PCAWG-OV samples using the method described in Menghi et al[21]. The number of duplication
705 events per chromosome normalized by chromosome length per sample was used to calculate a
706 score relative to the expected number of duplication events per chromosome per sample. The
707 scores ranged from -1.11 to 0.53 with an average score of 0.02.
708
709 *Mutational signatures.* Motif matrices were extracted using the SomaticSignatures R package[59]
710 and the weights of all known COSMIC signatures were determined using the deconstructSigs R
711 package[60] for 44 deep WGS BriTROC-1 samples and 109 PCAWG-OV samples. SNV signatures
712 showing an exposure >0 for at least one sample were retained. The rcorr function in the Hmisc R
713 package[61] was used to calculate the correlation matrix between the remaining SNV and CN
714 signature exposures.
715
716 The significance of all observed correlations was estimated from a t-distribution where the null
717 hypothesis was that the true correlation was 0. All reported p-values have been adjusted for
718 multiple testing with Benjamini & Hochberg (BH) method[62]. Comparison plots can be found in
719 Supplementary Figure 6.
720
721 *Mutated pathways:* A combined set of 479 samples (44 deep WGS BriTROC-1, 112 PCAWG-OV
722 and 323 TCGA) showing at least one driver mutation was used for mutated pathway enrichment
723 analysis. We focused on 765 driver genes reported by Cancer Genome Interpreter (CGI)[63]. SNVs,
724 INDELs, amplifications (CN>5) or deletions (CN<0.4) affecting these genes were considered *bona*
725 *fide* driver mutations if CGI predicted them as TIER1 or TIER2 (Supplementary Tables 4 and 5,

726 see URL 2, run date: 2018-01-13). 320 of the 765 genes were mutated in a least one case. These
727 genes were used to test for enriched pathways in the Reactome database using the ReactomePA
728 R package[64] with a p-value cutoff of 0.05 and q-value cutoff of 0.05. Pathways mutated in at least
729 5% of the cohort (n≥24) were retained. For each pathway, patients were split into two groups:
730 those with mutated genes in the pathways, and those with wild-type genes in the pathways. A one-
731 sided Mann-Whitney was carried out for each signature to determine if the exposure was
732 significantly higher in mutated cases versus wild-type cases. After multiple testing correction using
733 the Benjamini & Hochberg method (thresholding the p-value <0.005 and the median difference in
734 exposures ≥0.1), 186 pathways were significantly enriched. Visual inspection revealed significant
735 redundancy in the list and 9 representative pathways were manually selected as a final output
736 (Supplementary Table 6).
737
738 *Mutated genes:* A combined set of 479 samples (44 deep WGS BriTROC-1, 112 PCAWG-OV and
739 323 TCGA) was used test if signature exposures were significantly higher in cases with mutated
740 driver genes, including *NF1*, *PTEN*, *BRCA1*, *BRCA2*, *PIK3CA*, *MYC* and *CDK12*. Patients were
741 split into two groups: those with the mutated gene and those with wild-type genes. A one-sided
742 Mann-Whitney was carried out for each signature to determine if the exposure was significantly
743 higher in mutated cases versus wild-type cases. After multiple testing correction using the
744 Benjamini & Hochberg method (thresholding the p-value <0.05 and the median difference in
745 exposures ≥0.0.08), 10 gene/signature combinations were significantly enriched (Supplementary
746 Table 6).

747 Survival analysis

748 *Censoring and truncation:* Overall survival in BriTROC-1 patients was calculated from the date of
749 enrolment to the date of death or the last documented clinical assessment, with data cutoff at 1
750 December 2016. As the BriTROC-1 study only enrolled patients with relapsed disease, left
751 truncation was used in the survival analysis. In addition, cases where the patient was not
752 deceased were right censored. Survival data for the PCAWG-OV and TCGA cohorts were right
753 censored as required (left truncation was not necessary). The combined samples were split into
754 training (100% BriTROC-1, 70% PCAWG-OV and 70% TCGA = 417) and test (30% PCAWG-OV

and 30% TCGA = 158) cohorts. All of the BriTROC-1 samples were used in the training set to avoid issues calculating prediction performance on left-truncated data.

Cox regression: As the signature exposures for a given sample summed to 1, it was necessary to select one normalizing signature to perform regression. Signature 5 was chosen as it showed the lowest variability across the cohorts. To avoid division errors all 0 signature exposures were converted to 0.02. The remaining signature exposures were normalized taking the log ratio of their exposure to signature 5's exposure. A Cox proportional hazards model was fitted on the training set, with the signature exposures as covariates, stratified by cohort (BriTROC-1, PCAWG-OV:AU, PCAWG-OV:US, TCGA) and age (<39; 40:44; 45:49; 50:54; 55:59; 60:64; 65:69; 70:74; 75:79; >80), using the survival package in Bioconductor[65]. After fitting, the model was used to predict risk in the test set and performance was assessed using the concordance index calculation in the survcomp package in Bioconductor[47]. A final Cox regression was performed using all data for reporting of hazard ratios and p-values.

Unsupervised clustering of patients using signature exposures

Hierarchical clustering of the exposure vectors of the 575 samples used in the survival analysis was performed using the NbClust[66] package in R. The optimal number of clusters was 3 as determined by a consensus voting approach across 23 metrics for choosing the optimal numbers of clusters. 12/23 metrics reported 3 clusters as the optimal number. A Cox proportional hazards model was fitted using the cluster labels as covariates, stratified by cohort (BriTROC-1, PCAWG-OV:AU, PCAWG-OV:US, TCGA) and age (<39; 40:44; 45:49; 50:54; 55:59; 60:64; 65:69; 70:74; 75:79; >80), using the survival package in Bioconductor[65].

Analysis of copy-number signature changes during treatment

Thirty-six BriTROC-1 cases with matched diagnosis and relapse samples were used to investigate the effects of treatment on signature exposures. A linear model was fitted to test for treatment effects with exposure at relapse as the dependent variable and exposure at diagnosis, age at diagnosis, number of lines of chemotherapy, and days between diagnosis and relapse as independent variables. Prior to fitting, age at diagnosis was centered and exposures transformed by *log(x+0.1)* to ensure normality. Fitting was done using the *lm()* function in R.

To test whether signature exposures at diagnosis were predictive of platinum sensitivity, a generalized linear model with Binomial error was fitted using type of relapse (platinum-sensitive or platinum-resistant) as the dependent variable and exposure at diagnosis and age at diagnosis as independent variables.

789 ## Data Availability

790 Sequence data that support the findings of this study have been deposited in the European
791 Genome-phenome Archive with the accession code EGAS00001002557. All code required to
792 reproduce the analysis outlined in this manuscript can be found in the following repository (see
793 URL 3).
794

795 ## URLs

796 1. https://gmacintyre.shinyapps.io/sWGS_power/
797 2. https://www.cancergenomeinterpreter.org/home
798 3. https://bitbucket.org/britroc/cnsignatures
799
800

Methods-only References

41. Piskorz, A.M. *et al.* Methanol-based fixation is superior to buffered formalin for next-generation sequencing of DNA from clinical cancer samples. *Ann Oncol* **27**, 532-539 (2016).
42. Scheinin, I. *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 Res* **24**, 2022-32 (2014).
43. Macintyre, G., Ylstra, B. & Brenton, J.D. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends Genet* **32**, 530-42 (2016).
44. bcbio-nextgen. (2017).
45. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108 (2016).
46. Koboldt, D.C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-5 (2009).
47. Garrison E, M.G. Haplotype-based variant detection from short-read sequencing. in *arXiv* (2012).
48. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15 10 1-15 10 18 (2016).
49. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-21 (2012).
50. Schumacher, S. pancan12_absolute.segtab.txt. (2015).
51. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).
52. Al-Kateb, H., Nguyen, T.T., Steger-May, K. & Pfeifer, J.D. Identification of major factors associated with failed clinical molecular oncology testing performed by next generation sequencing (NGS). *Mol Oncol* **9**, 1737-43 (2015).
53. Grün, B. & Leisch, F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J Stat Soft* **28**, 35 (2008).
54. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
55. Brunet, J.-P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-4169 (2004).
56. Huebschmann, D., Gu, Z. & Schlesner, M. YAPSA: Yet Another Package for Signature Analysis. R package version 1.2.0. (2015).
57. Farmery, J.H.S., Mike L; Lynch Andy G. Telomerecat: A Ploidy-Agnostic Method For Estimating Telomere Length From Whole Genome Sequencing Data. in *bioRxiv* (2017).
58. Kim, S. ppcor: Partial and Semi-Partial (Part) Correlation. (2015).
59. Gehring, J.S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673-5 (2015).
60. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
61. Harrell, F.E. Hmisc: Harrell Miscellaneous. R package version 4.0-0. (2016).
62. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
63. Tamborero, D. *et al.* Cancer Genome Interpreter Annotates The Biological And Clinical Relevance Of Tumor Alterations. in *bioRxiv* (2017).
64. Yu, G. & He, Q.Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477-9 (2016).
65. Therneau, T.M., Grambsch, Patricia M. *Modeling Survival Data: Extending the Cox Model*, (Springer, New York, 2000).
66. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Soft* **61**, 36 (2014).

856
857
858
859
860
861
862

a

Compute absolute CN from shallow WGS

Absolute copy-number

Chromosomes

Derive CN feature distributions

Breakpoint count per 10MB

Copy-number

Copy-number changepoint

Breakpoint count per chr arm

Length of chains of oscilating copy-number

Segment size

Fit optimal number of mixture model components

N=3

N=8

N=7

N=5

N=3

N=10

Compile sum-of-posteriors matrix

Tumor by component matrix

Tumors (N=117)

Components (N=36)

Perform non-negative matrix factorisation

Tumor by signature matrix

Tumors (N=117)

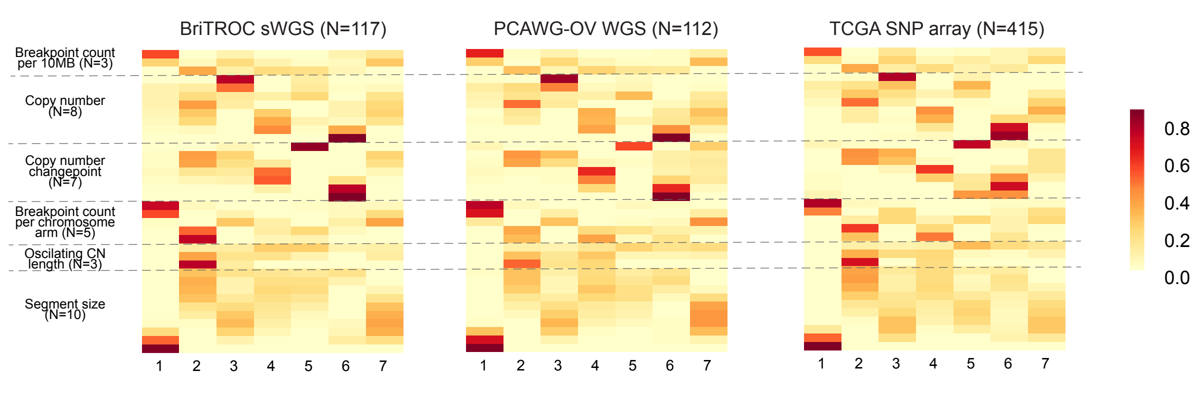
CN signatures

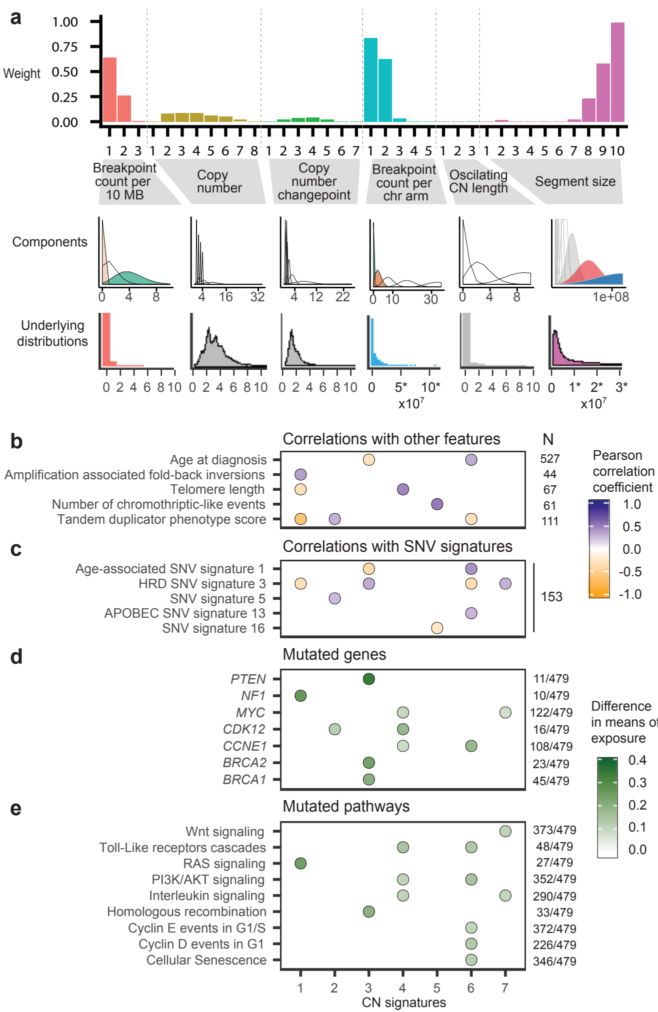
Signature by component matrix

CN signatures

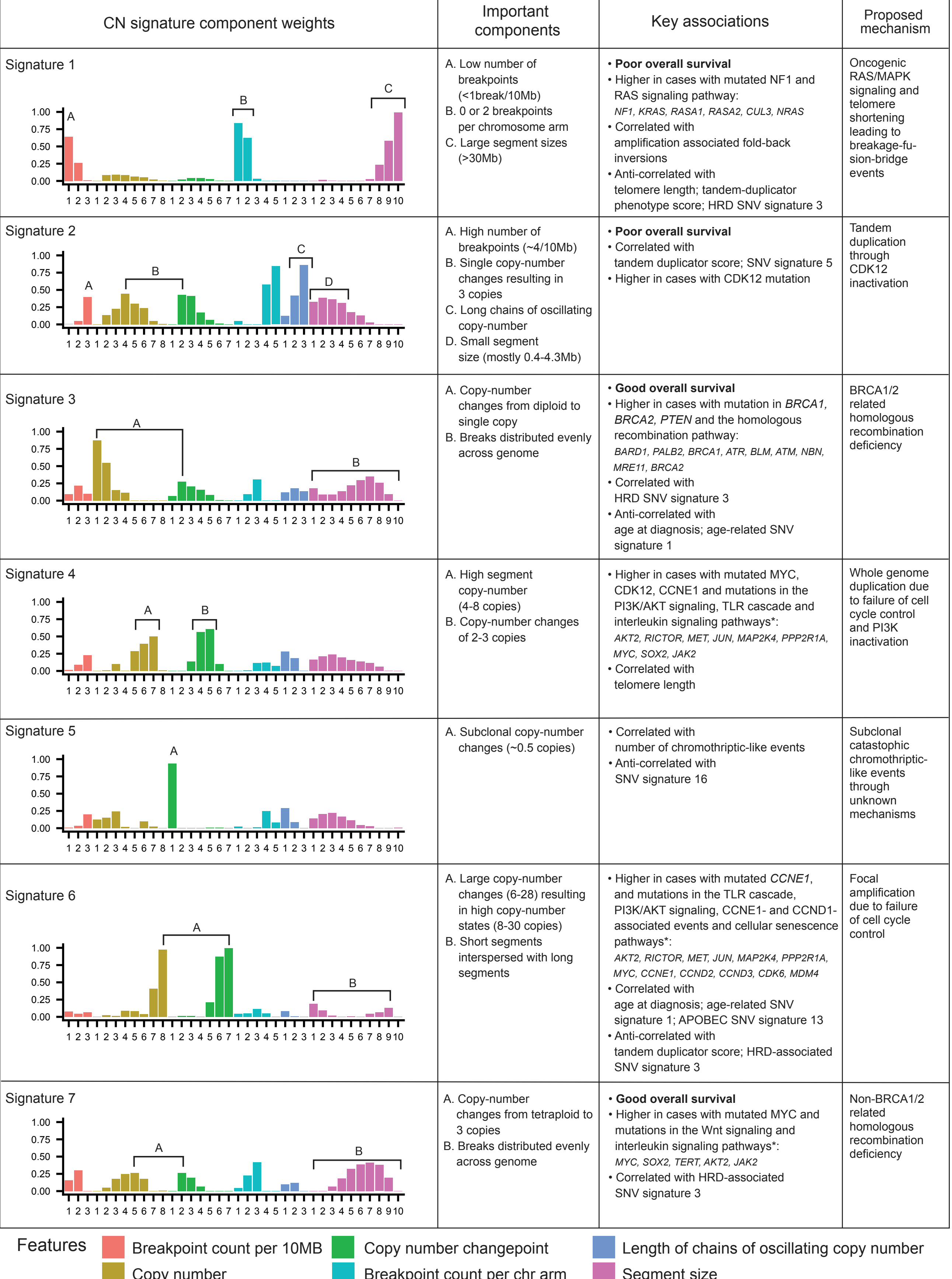
Components (N=36)

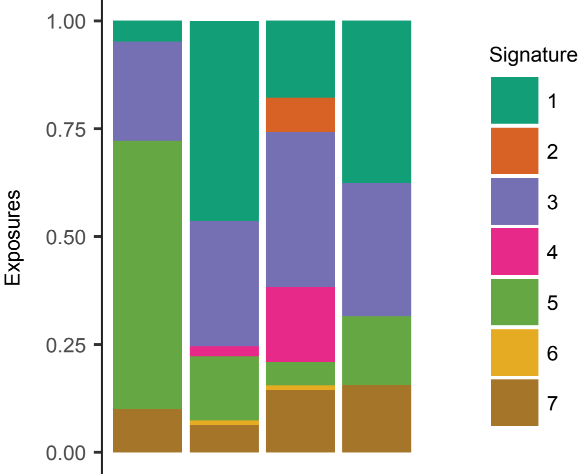
b

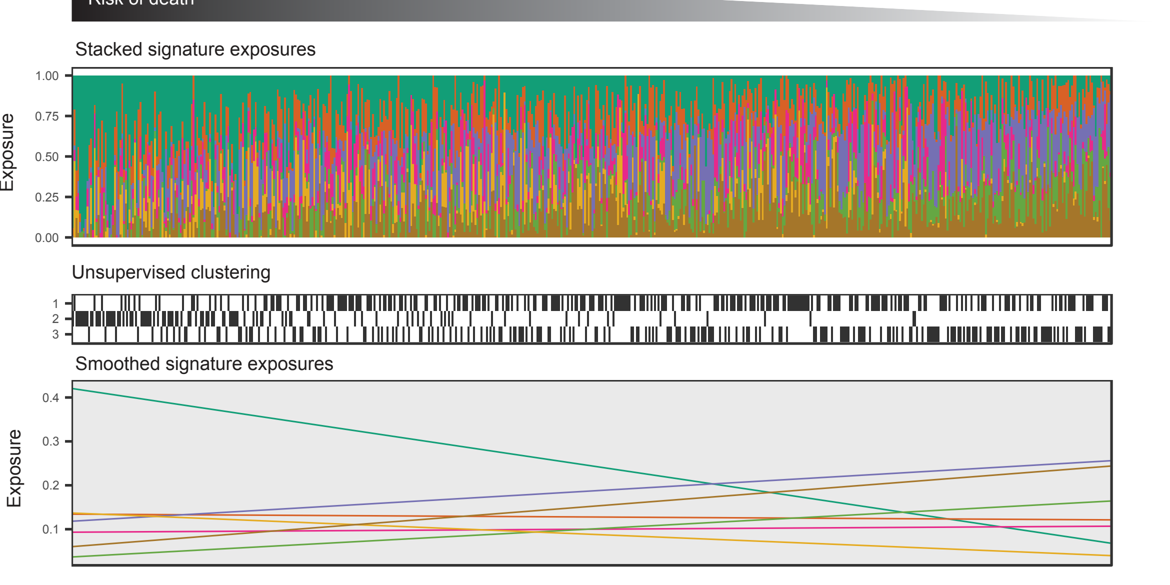




a
Weight
1.00
0.75
0.50
0.25
0.00
Breakpoint count per 10 MB
Copy number
Copy number changepoint
Breakpoint count per chr arm
Oscilating CN length
Segment size
Components
Underlying distributions
x10^7
b
Correlations with other features
N
Age at diagnosis
527
Amplification associated fold-back inversions
44
Telomere length
67
Number of chromothriptic-like events
61
Tandem duplicator phenotype score
111
Pearson correlation coefficient
c
Correlations with SNV signatures
Age-associated SNV signature 1
HRD SNV signature 3
SNV signature 5
APOBEC SNV signature 13
SNV signature 16
153
d
Mutated genes
PTEN
11/479
NF1
10/479
MYC
122/479
CDK12
16/479
CCNE1
108/479
BRCA2
23/479
BRCA1
45/479
Difference in means of exposure
e
Mutated pathways
Wnt signaling
373/479
Toll-Like receptors cascades
48/479
RAS signaling
27/479
PI3K/AKT signaling
352/479
Interleukin signaling
290/479
Homologous recombination
33/479
Cyclin E events in G1/S
372/479
Cyclin D events in G1
226/479
Cellular Senescence
346/479
CN signatures


CN signature component weights
Important components
Key associations
Proposed mechanism
Signature 1
1.00 0.75 0.50 0.25 0.00
A B C
1 2 3 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 1 2 3 4 5 1 2 3 1 2 3 4 5 6 7 8 9 10
A. Low number of breakpoints (<1break/10Mb)
B. 0 or 2 breakpoints per chromosome arm
C. Large segment sizes (>30Mb)
• Poor overall survival
• Higher in cases with mutated NF1 and RAS signaling pathway:
NF1, KRAS, RASA1, RASA2, CUL3, NRAS
• Correlated with amplification associated fold-back inversions
• Anti-correlated with telomere length; tandem-duplicator phenotype score; HRD SNV signature 3
Oncogenic RAS/MAPK signaling and telomere shortening leading to breakage-fusion-bridge events
Signature 2
A B C D
A. High number of breakpoints (~4/10Mb)
B. Single copy-number changes resulting in 3 copies
C. Long chains of oscillating copy-number
D. Small segment size (mostly 0.4-4.3Mb)
• Poor overall survival
• Correlated with tandem duplicator score; SNV signature 5
• Higher in cases with CDK12 mutation
Tandem duplication through CDK12 inactivation
Signature 3
A B
A. Copy-number changes from diploid to single copy
B. Breaks distributed evenly across genome
• Good overall survival
• Higher in cases with mutation in BRCA1, BRCA2, PTEN and the homologous recombination pathway:
BARD1, PALB2, BRCA1, ATR, BLM, ATM, NBN, MRE11, BRCA2
• Correlated with HRD SNV signature 3
• Anti-correlated with age at diagnosis; age-related SNV signature 1
BRCA1/2 related homologous recombination deficiency
Signature 4
A B
A. High segment copy-number (4-8 copies)
B. Copy-number changes of 2-3 copies
• Higher in cases with mutated MYC, CDK12, CCNE1 and mutations in the PI3K/AKT signaling, TLR cascade and interleukin signaling pathways*:
AKT2, RICTOR, MET, JUN, MAP2K4, PPP2R1A, MYC, SOX2, JAK2
• Correlated with telomere length
Whole genome duplication due to failure of cell cycle control and PI3K inactivation
Signature 5
A
A. Subclonal copy-number changes (~0.5 copies)
• Correlated with number of chromothriptic-like events
• Anti-correlated with SNV signature 16
Subclonal catastophic chromothriptic-like events through unknown mechanisms
Signature 6
A B
A. Large copy-number changes (6-28) resulting in high copy-number states (8-30 copies)
B. Short segments interspersed with long segments
• Higher in cases with mutated CCNE1, and mutations in the TLR cascade, PI3K/AKT signaling, CCNE1- and CCND1-associated events and cellular senescence pathways*:
AKT2, RICTOR, MET, JUN, MAP2K4, PPP2R1A, MYC, CCNE1, CCND2, CCND3, CDK6, MDM4
• Correlated with age at diagnosis; age-related SNV signature 1; APOBEC SNV signature 13
• Anti-correlated with tandem duplicator score; HRD-associated SNV signature 3
Focal amplification due to failure of cell cycle control
Signature 7
A B
A. Copy-number changes from tetraploid to 3 copies
B. Breaks distributed evenly across genome
• Good overall survival
• Higher in cases with mutated MYC and mutations in the Wnt signaling and interleukin signaling pathways*:
MYC, SOX2, TERT, AKT2, JAK2
• Correlated with HRD-associated SNV signature 3
Non-BRCA1/2 related homologous recombination deficiency
Features
Breakpoint count per 10MB
Copy number
Copy number changepoint
Breakpoint count per chr arm
Length of chains of oscillating copy number
Segment size


1.00
0.75
0.50
0.25
0.00
Exposures
BRCA2 germline mutation carriers + somatic LOH (n=4)
Signature
1
2
3
4
5
6
7



Risk of death
Stacked signature exposures
Exposure
1.00
0.75
0.50
0.25
0.00
Unsupervised clustering
1
2
3
Smoothed signature exposures
Exposure
0.4
0.3
0.2
0.1
Tumors ordered by decreasing risk of death (n=575)
CN signature
1
2
3
4
5
6
7