Article

A pan-cancer compendium of chromosomal instability

https://doi.org/10.1038/s41586-022-04789-9

Received: 31 July 2020

Accepted: 21 April 2022

Published online: 15 June 2022

Check for updates

Ruben M. Drews[1], Barbara Hernando[2], Maxime Tarabichi[3,4], Kerstin Haase[3,5], Tom Lesluyes[3], Philip S. Smith[1], Lena Morrill Gavarró[1], Dominique-Laurent Couturier[1,6], Lydia Liu[3,7], Michael Schneider[1], James D. Brenton[1,8,9], Peter Van Loo[3], Geoff Macintyre[1,2,10] ✉ & Florian Markowetz[1,10] ✉

Chromosomal instability (CIN) results in the accumulation of large-scale losses, gains and rearrangements of DNA[1]. The broad genomic complexity caused by CIN is a hallmark of cancer[2]; however, there is no systematic framework to measure different types of CIN and their effect on clinical phenotypes pan-cancer. Here we evaluate the extent, diversity and origin of CIN across 7,880 tumours representing 33 cancer types. We present a compendium of 17 copy number signatures that characterize specific types of CIN, with putative aetiologies supported by multiple independent data sources. The signatures predict drug response and identify new drug targets. Our framework refines the understanding of impaired homologous recombination, which is one of the most therapeutically targetable types of CIN. Our results illuminate a fundamental structure underlying genomic complexity in human cancers and provide a resource to guide future CIN research.

CIN has complex consequences, including loss or amplification of driver genes, focal rearrangements, extrachromosomal DNA, micronuclei formation and activation of innate immune signalling[1]. This leads to associations with disease stage, metastasis, poor prognosis and therapeutic resistance[3]. The causes of CIN are also diverse and include mitotic errors, replication stress, homologous recombination deficiency (HRD), telomere crisis and breakage fusion bridge cycles, among others[1,4].

Because of the diversity of these causes and consequences, CIN is generally used as an umbrella term. Measures of CIN either divide tumours into broad categories of high or low CIN[5], are restricted to a single aetiology such as HRD[6], are limited to a particular genomic feature such as whole-chromosome-arm changes[7], or can only be quantified in specific cancer types[8,9]. As a result, there is no systematic framework to comprehensively characterize the diversity, extent and origins of CIN pan-cancer, or to define how different types of CIN within a tumour relate to clinical phenotypes. Here we present a robust analysis framework to quantitatively measure different types of CIN across cancer types.

Deconstructing CIN

We derived 7,880 high-quality absolute copy number profiles across 33 tumour types using single-nucleotide polymorphism (SNP) array data from The Cancer Genome Atlas (TCGA) (Extended Data Fig. 1a). Extending our previously developed framework for quantifying signatures of CIN in ovarian cancer[8], we determined that 6,335 of the 7,880 samples (80%) had detectable CIN and were suitable for pan-cancer detection of copy number signatures (Extended Data Fig. 1b). This estimate was consistent with previous pan-cancer estimates of CIN[10] (Extended Data Fig. 1c–e).

Using these 6,335 genome-wide copy number profiles, we computed distributions of five fundamental copy number features previously demonstrated to encode patterns of copy number changes that represent different underlying causes of CIN[8] (Extended Data Fig. 2a and Supplementary Methods). These features included: the copy number change between a segment and the neighbouring segment; segment length; breakpoint count per 10 Mb; breakpoint count per chromosome arm; and length of chains of oscillating copy number states. Only segments that deviated from a normal, diploid state were considered for the segment size and changepoint features. We did not include a feature representing the copy number of a segment to avoid redundant signatures that encode the same aetiology across different ploidy backgrounds.

We applied mixture modelling to define distinct components for each cohort-wide feature distribution, identifying a total of 43 mixture components across the 5 features (Extended Data Fig. 2b, c and Supplementary Methods). Conceptually, these components represent the basic building blocks for defining CIN processes. We used these mixture components to encode each tumour genome by probabilistically assigning copy number events to these components, resulting in a 6,335 × 43 dimensional matrix. We then applied a Bayesian implementation of non-negative matrix factorization to identify copy number signatures (Extended Data Figs. 2d and 3a, b). We first used the complete matrix and found 10 pan-cancer copy number signatures, then used subsets of the matrix representing individual cancer types with

[1]Cancer Research UK Cambridge Institute, Cambridge, UK. [2]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [3]The Francis Crick Institute, London, UK. [4]Institute for Interdisciplinary Research (IRIBHM), Université Libre de Bruxelles, Brussels, Belgium. [5]Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany. [6]Medical Research Council Biostatistics Unit, Cambridge, UK. [7]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [8]Addenbrooke's Hospital, Cambridge, UK. [9]Department of Oncology, University of Cambridge, Cambridge, UK. [10]These authors contributed equally: Geoff Macintyre, Florian Markowetz. ✉e-mail: gmacintyre@cnio.es; florian.markowetz@cruk.cam.ac.uk

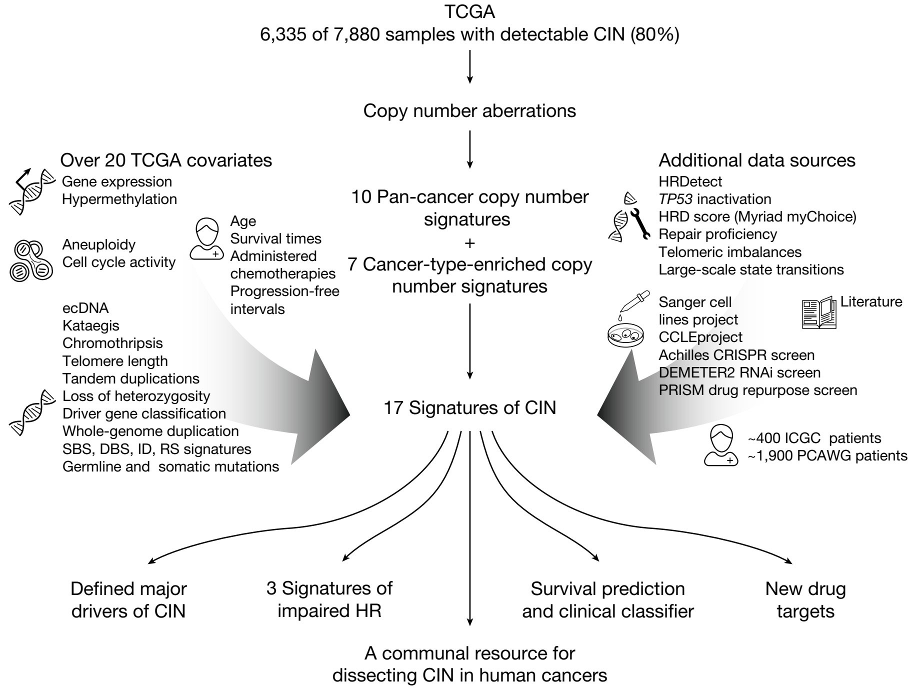

TCGA
6,335 of 7,880 samples with detectable CIN (80%)
Copy number aberrations
Over 20 TCGA covariates
Gene expression
Hypermethylation
Aneuploidy
Cell cycle activity
Age
Survival times
Administered chemotherapies
Progression-free intervals
ecDNA
Kataegis
Chromothripsis
Telomere length
Tandem duplications
Loss of heterozygosity
Driver gene classification
Whole-genome duplication
SBS, DBS, ID, RS signatures
Germline and somatic mutations
10 Pan-cancer copy number signatures
+
7 Cancer-type-enriched copy number signatures
17 Signatures of CIN
Additional data sources
HRDetect
TP53 inactivation
HRD score (Myriad myChoice)
Repair proficiency
Telomeric imbalances
Large-scale state transitions
Sanger cell lines project
CCLEproject
Achilles CRISPR screen
DEMETER2 RNAi screen
PRISM drug repurpose screen
Literature
~400 ICGC patients
~1,900 PCAWG patients
Defined major drivers of CIN
3 Signatures of impaired HR
Survival prediction and clinical classifier
New drug targets
A communal resource for dissecting CIN in human cancers

Fig. 1 | Study overview. This schematic summarizes our robust analysis framework, which uses copy number to derive pan-cancer copy number signatures and provide insights. On the left and right are lists of the datasets used to support the signature aetiologies and insights. CCLE, Cancer Cell Line Encyclopedia; DBS, doublet-base substitution; ecDNA, extrachromosomal DNA; RS, rearrangement signature.

at least 100 samples, and found an additional 7 signatures (Extended Data Fig. 3b–e and Supplementary Methods). We merged both sets of signatures and computed their activities using linear combination decomposition to yield a pan-cancer compendium of 17 copy number signatures and their activities in tumours across the 33 cancer types (Extended Data Figs. 3f, g and 4 and Supplementary Figs. 1 and 2).

We validated this approach by correctly identifying signatures in a collection of simulated cancer genomes with copy number changes caused by five well-studied mutational processes (Supplementary Figs. 3–6 and Supplementary Methods). We used a second simulation study to derive signature-specific activity thresholds, to test the stability of signature definitions and to test the stability of signature activities (Methods, Extended Data Fig. 5 and Supplementary Fig. 7). We then tested the robustness of our approach across different high-throughput technologies comparing signature definitions and activities across five platforms: SNP 6.0 without matched normal, whole-genome sequencing (WGS) downsampled to SNP 6.0 positions, WGS downsampled to shallow WGS, on-target whole-exome sequencing (WES) and off-target WES. Quantification of signature activity was robust across all platforms. Signature identification was possible across the WGS platforms but performance deteriorated for WES (Extended Data Fig. 6).

Putative causes underlying each signature

To determine the putative causes underlying each of the 17 signatures (named CX1 to CX17), we developed a data integration framework and assigned a confidence score to each signature aetiology based on the quality and extent of supporting data (Extended Data Fig. 7). To propose putative aetiologies, we used the patterns of copy number change encoded by the signature (Extended Data Fig. 4, Supplementary Figs. 8 and 9 and Methods) and signature associations with known cancer driver mutations (Extended Data Fig. 8a and Supplementary Figs. 10–17). We used these driver gene associations as markers for putative pathways involved in the aetiologies and assumed the same pathway deregulation for samples where no driver gene was mutated (similar to how BRCAness is defined in the absence of *BRCA1* or *BRCA2* mutation[11]). In many cases, the signature pattern was already suggestive of a mechanism (for example, whole-chromosome missegregation). Once a putative cause was proposed, we sought additional supporting data (Fig. 1, Extended Data Figs. 8 and 9 and Supplementary Methods) including: data from two additional patient cohorts and their clinical metadata (approximately 1,900 patients from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project and approximately 400 patients from the International Cancer Genome Consortium (ICGC) project); five types of mutational signatures (single-base substitution (SBS), insertion–deletion (ID), doublet base substitutions, ovarian copy number and rearrangement); 14 molecular features (somatic point mutations, gene expression, cell cycle score, aneuploidy score, whole-chromosome copy number aberrations (CNAs), tandem duplications, loss of heterozygosity, chromothripsis, kataegis, whole-genome duplication status, telomere length and elongation machinery activity, extrachromosomal DNA and centrosome amplification score (CA20)), and 11 DNA repair-specific features (germline *BRCA1/2* mutations, *BRCA1* and *RAD51C* hypermethylation data, HRDetect response, HRD score (Myriad myChoice), *TP53* inactivation score, telomeric imbalances score, large-scale state transition score, loss of heterozygosity score, DNA repair proficiency score, protein expression score for 23 DNA-damage repair genes and PCAWG structural variants with associated microhomologies). Here we provide a synthesis of the data supporting the putative aetiologies (summarized in Fig. 2).

Mitotic signatures

CX1, CX6 and CX14 all encoded patterns related to whole-arm or whole-chromosome changes and significantly correlated with direct counts of whole-chromosome changes (Supplementary Fig. 18). This suggested putative causes resulting in chromosome missegregation during mitosis. In agreement with this hypothesis, CX14 had significantly higher activity in tumours with inactivating mutations in *CIC*[12]; CX1 with mutations in *CIC*[12], *VHL*[13] and *PBRM1* (ref. [14]); and CX6 with mutations in *CUL1* (ref. [15]) and *RAC1* (ref. [16]) (Extended Data Fig. 8a). Each of the three signatures correlated with downregulation of telomerase activity (Supplementary Fig. 19b), with CX1 also being negatively

Signature	Pan-cancer prevalence (%)	Putative cause	Confidence	Pattern of change
CX1	80	Chromosome missegregation via defective mitosis and/or telomere dysfunction	★★★	Whole-arm/chromosome changes
CX2	69	IHR	★★★	Short-to-medium-sized, oscillating, single-copy changes
CX3	63	IHR with replication stress and impaired damage sensing	★★★	Long-sized, single-copy changes
CX4	58	PI3K–AKT-mediated toleration of whole-genome duplication	★★★	Medium-sized, clustered and oscillating, two-copy changes
CX5	56	IHR with replication stress	★★★	Medium-sized, two-to-three-copy changes
CX6	51	Chromosome missegregation via defective mitosis	★★★	Whole-arm/chromosome changes
CX7	50	Unknown	☆☆☆	Long-sized, less than one-copy change
CX8	38	Replication stress	★★★	Short-to-medium-sized, oscillating, low-level amplifications
CX9	32	Replication stress	★☆☆	Short-to-medium-sized, clustered, mid-level amplifications
CX10	29	Impaired NHEJ with replication stress	★★☆	Short-to-medium-sized, clustered and oscillating, two-copy changes
CX11	28	Replication stress	★★☆	Short-sized, clustered and oscillating, mid-level amplifications
CX12	27	Unknown	☆☆☆	Very short oscillating changes
CX13	26	Replication stress	★☆☆	Short-to-medium-sized, clustered, high-level amplifications
CX14	26	Chromosome missegregation via defective mitosis	★★☆	Whole-arm/chromosome, less than one-copy changes
CX15	25	Unknown	☆☆☆	Long-sized, less than one-copy changes
CX16	17	Unknown	★☆☆	Very short, clustered and oscillating, two-to-three-copy changes
CX17	16	Unknown	★★☆	Short-to-long-sized, one-to -three-copy changes

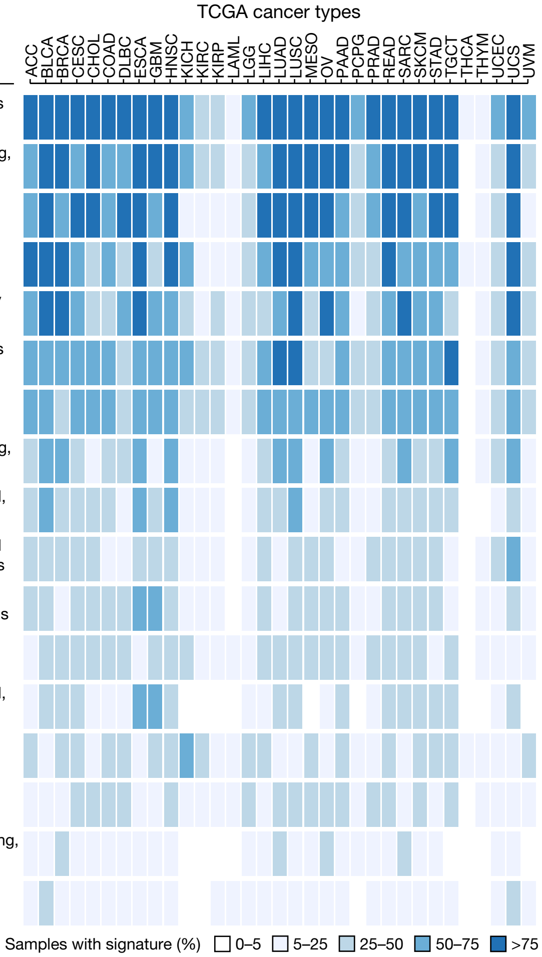


Fig. 2 | Proposed aetiologies and prevalence of copy number signatures. A summary of the pan-cancer frequency, proposed aetiology (where possible), aetiology confidence rating, pattern of copy number change and distribution across cancer types is provided for each signature. Signatures are labelled on the basis of pan-cancer prevalence, with signature CX1 having the highest pan-cancer frequency. Confidence measures for each signature aetiology are indicated by a star rating. The heatmap shows the signature frequency for each of the 33 cancer types. NHEJ, non-homologous end joining; IHR, impaired homologous recombination; ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, oesophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LAML, acute myeloid leukemia; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; OV, ovarian serous cystadenocarcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumours; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

correlated with telomere length (Supplementary Fig. 19a) and associated with a lack of *TERC* and *TERT* amplification and expression (Supplementary Fig. 19c–e, and 20). Therefore, telomere shortening may have a key role in the mechanisms underlying these signatures[4]. CX1 positively correlated with the 'clock-like' SBS1 signature, suggesting that these errors might also be mediated via a natural ageing process such as age-related telomere attrition[4] (Extended Data Fig. 8 and Supplementary Fig. 21).

Signatures of impaired homologous recombination

CX2, CX3 and CX5 all exhibited patterns that had previously been shown to associate with impaired homologous recombination (IHR): CX2 showed a pattern of short-to-medium-sized, oscillating changes associated with tandem duplications[17]; CX5 showed medium-sized events associated with tandem duplication[17]; and CX3 showed long-sized, single-copy changes with associated loss of heterozygosity[18,19] (Extended Data Fig. 4 and Supplementary Figs. 18 and 22). All three signatures were observed at significantly higher levels in tumours with somatic *BRCA1* mutation, independently of each other (Extended Data Figs. 8a and 9a and Supplementary Table 12). This suggested varying roles for disruption of HR as underlying causes[11]. Several lines of evidence supported the link between these signatures and HR: increased CX2, CX3 and CX5 activity across germline-mutated *BRCA1* carriers (and *BRCA2* carriers for CX3); higher activity in cases with methylated *RAD51C* (except CX5)[20] (Extended Data Fig. 9a); correlation with tandem duplication scores[17] (Supplementary Fig. 22), rearrangement signatures 1, 3 and 5 (ref. [21]) (Supplementary Fig. 23), SBS3 signature and ID6 (Supplementary Fig. 21), centrosome amplification score[22] (Supplementary Fig. 24), and ovarian copy number signatures 3 and 7 (ref. [8]) (Supplementary Fig. 25); association with loss of heterozygosity[18], chromothripsis[23] (except CX3) and kataegis[24](Supplementary Fig. 18); increased utilization of theta-mediated end joining and single-strand annealing backup repair pathways visible as microhomologies at breakpoints[11] (Supplementary Fig. 26); as well as correlation with seven HRD metrics[25]

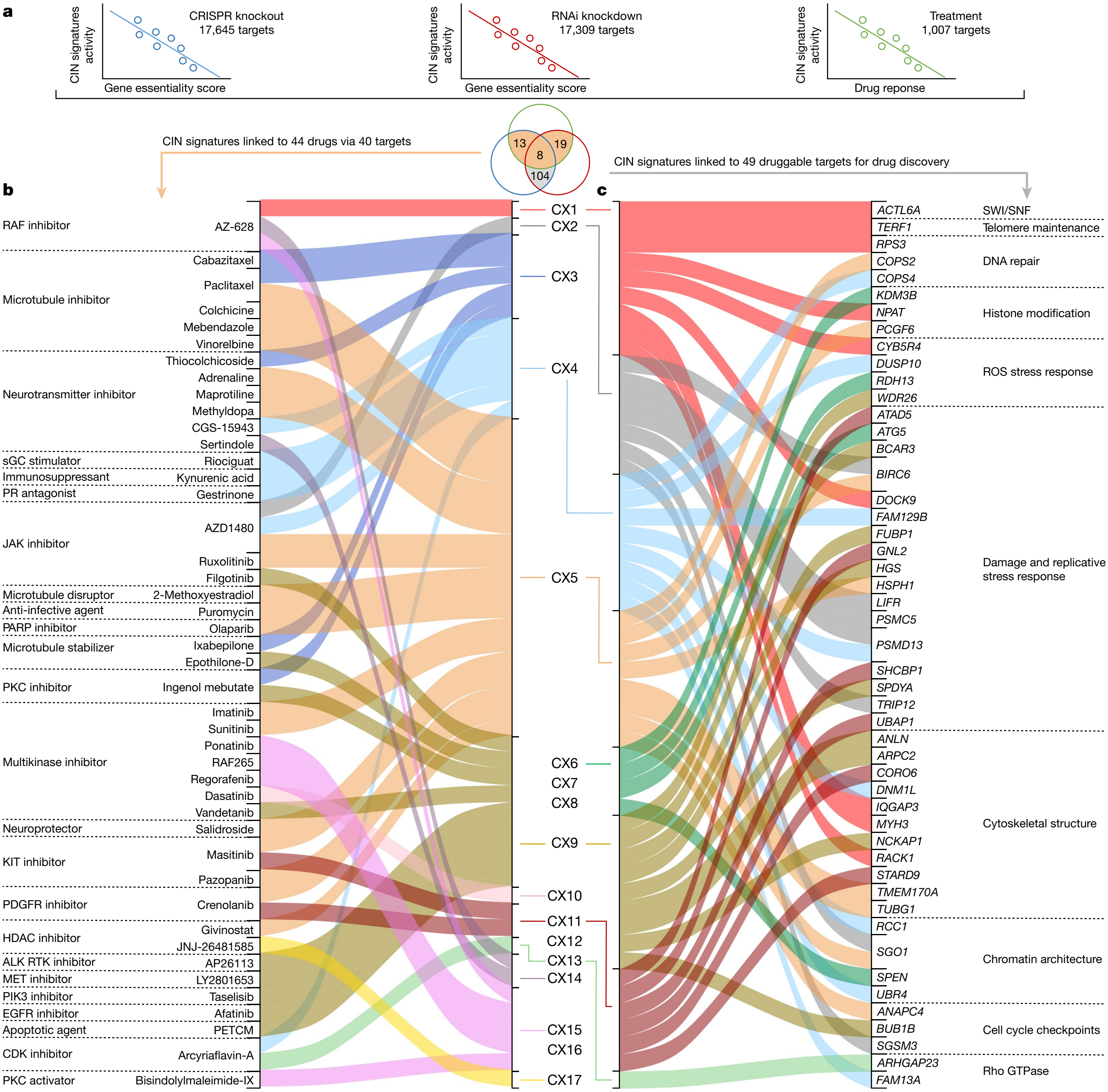


Fig. 3 | Signatures as biomarkers for drug response and discovery of novel drug targets. a, A schematic showing how response biomarkers and novel drug targets were found by correlating signature activities with gene essentiality determined by CRISPR–Cas9 or RNAi screens, and with response to drug perturbations measured as the area under the dose response curve, across 297 cell lines. The Venn diagram shows the overlap of significant correlations for each of the signature to target gene associations. The colour of the circles in the Venn diagram matches the schematic above, and the shaded areas indicate which results relate to **b** and **c**. **b**, A summary of the significant associations between copy number signatures and drug response to 44 therapies. Each signature on the right is linked to a therapy on the left if the signature is predictive of response to CRISPR and/or RNAi perturbation of a target gene, and treatment with a therapy that targets that gene. HDAC, histone deacetylase; PR, progesterone receptor; sGC, soluble guanylate cyclase. **c**, A summary of the significant associations between copy number signatures and target gene perturbation. Each signature on the left is linked to a target gene on the right if the signature is predictive of response to CRISPR and RNAi perturbation of the target gene. The listed targets were filtered for druggability according to their structure or by ligand-based approaches ($n = 104$) and their previous known association with CIN ($n = 49$). ROS, reactive oxygen species; IHR, impaired homologous recombination.

(Extended Data Fig. 9d). The strength of these associations increased from CX2 to CX5 and to CX3. This suggested an increasing spectrum of CIN complexity associated with disruptions in HR-mediated repair. Indeed, CX2 appears to be only associated with disruption of HR, whereas CX5 and CX3 have associations that indicate the involvement of replication stress (via amplification and overexpression of *MAPK1* (ref. [26]), *PPP2R1A*[27] and *U2AF1* (ref. [28])). The larger copy number changes observed for CX5 and CX3 suggest faster cell cycling and breaks carried

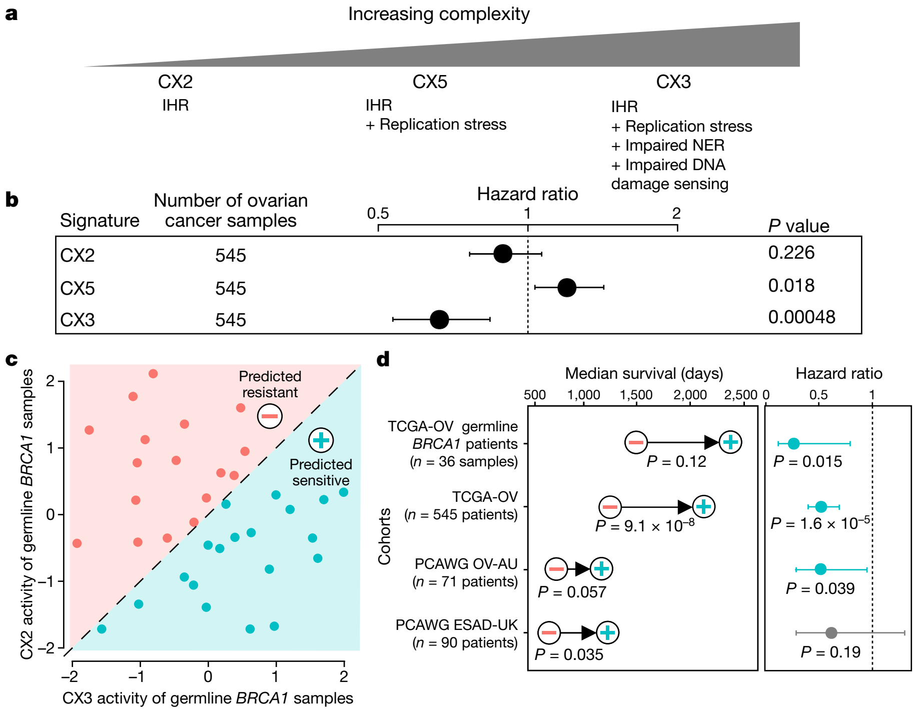


Fig. 4 | Predicting platinum sensitivity using IHR signatures. a, A proposed model of increasing CIN complexity for IHR signatures based on the signature aetiologies. **b**, Results for each IHR signature after training a Cox proportional hazards model to predict overall survival across 545 ovarian cancers treated with platinum-based chemotherapy. Hazard ratios, their 95% confidence interval and Wald test significance are reported. The dashed line indicates a hazard ratio of 1. **c**, A schematic of the clinical classifier built on CX3 and CX2 activities of ovarian cancer samples with germline *BRCA1* mutations. **d**, Results of survival analyses after applying the classifier from **c** to assign patients into predicted sensitive (plus symbol) or predicted resistant (minus symbol) groups. Each row displays results for each of the four cancer cohorts from the TCGA and PCAWG projects. Differences in median survival are indicated by the arrow, with *P* values from a log-rank test appearing below (Kaplan–Meier survival analysis). Hazard ratios and their 95% confidence interval of the predicted sensitive group compared to the predicted resistant group are obtained from Cox proportional hazards models correcting for stage and age of patients. The *P* value represents the corresponding Wald test. AU, Australian Project at the ICGC/PCAWG consortium; OV, ovarian cancer; ESAD, oesophageal cancer; UK, British project at the ICGC/PCAWG consortium.

through to mitosis[11], which was supported by strong correlation with cell cycle scores (Extended Data Fig. 9b) and increased CNAs estimated to occur during mitosis (Supplementary Fig. 27 and Supplementary Methods). Further associations were observed for CX3, including missense mutations in *ERCC2* (ref. [29]) and downregulation of key nucleotide excision repair (NER) genes suggesting defects in NER (Extended Data Fig. 9c and Supplementary Fig. 28), as well as *TP53* mutation suggesting impaired damage sensing[30] (Extended Data Fig. 8a). These CX3 associations are reminiscent of what has been termed BRCAness or HRD[11]. However, CX5, and especially CX2, appear to represent a more moderate impairment of HR. Therefore, we use the term IHR for the aetiology underlying all three signatures rather than HRD.

Whole-genome duplication signature

CX4 encompassed a unique pattern of copy number change with neighbouring segments separated by two copy changes (Extended Data Fig. 4), a pattern commonly used to define the presence of a whole-genome duplication (WGD) event[31]. CX4 was also associated with whole-chromosome changes (Extended Data Fig. 8b), a feature commonly observed in tetraploid cells due to increased mitotic errors[32]. The specific cause of WGD (endoreduplication, errors in cytokinesis or cell fusion[33]) was not evident from our data; however, this signature had high activity in tumours with *PIK3R2*, *AKT1* and *MAPK1* mutations, suggesting that tolerance to WGD may be mediated by PI3K–AKT activation[34,35] (Extended Data Fig. 8a).

Signature of impaired non-homologous end joining

CX10 displayed a pattern of clustered and oscillating copy number changes (Extended Data Fig. 4). Its activity was significantly higher in tumours with inactivating mutations in *FBXW7* and correlated with *FBXW7*-mutant-mediated tandem duplication class 1/2 (Extended Data Fig. 8 and Supplementary Fig. 22), suggesting impaired non-homologous end joining[17,36] as a putative cause. A significant increase in the proportion of breakpoints with microhomologies in samples with this signature was indicative of a lack of blunt-end joining, which is a hallmark of non-homologous end joining (Supplementary Fig. 29a).

Signatures of amplification

CX8, CX9, CX11 and CX13 encoded patterns of low-level, mid-level, mid-level and high-level amplifications, respectively (Extended Data Fig. 4). Higher activity of CX8 in the context of amplification and overexpression of *U2AF1* (ref. [28]) and *MAPK1* (ref. [26]), and for CX9 *ERBB3* (ref. [37]) (Extended Data Fig. 8), suggested replication stress as a putative cause. All four signatures were associated with increased cell cycle score (Supplementary Fig. 30), reinforcing replication stress as a causal factor. In addition, CX8, CX9 and CX13 were associated with APOBEC mutagenesis (SBS2 and/or SBS13 signatures; Supplementary Fig. 21a), and CX9 and CX11 were associated with ID signatures 1 and 2 (ref. [38]) (Supplementary Fig. 21). CX9 copy number changes were not part of oscillating chains; however, the remaining amplification signatures were. CX13 was strongly associated with extrachromosomal DNA circularization and amplification events (Supplementary Fig. 31); however, the specific mechanism causing the extrachromosomal DNA was not evident.

Unknown aetiologies

CX7, CX12, CX15, CX16 and CX17 did not have patterns of copy number change or associations clearly indicative of a putative cause

(Extended Data Figs. 4 and 8a). Therefore, these signatures currently have unknown aetiologies.

Cross-signature observations

Many covariates demonstrated associations with multiple signatures. Chromothripsis was linked with seven different signatures (Extended Data Fig. 8), suggesting that many potential aetiologies underpin these complex rearrangements. Replication stress was associated with eight signatures, highlighting it as a major source of CIN (Fig. 2). Different signatures showed a bias for occurrence before WGD (CX1, CX2, CX7 and CX15) or after WGD (CX3, CX5, CX6, CX8, CX9, CX13 and CX17), demonstrating the importance of WGD events in modulating CIN (Extended Data Fig. 8b and Supplementary Fig. 18e, f). Finally, signatures of APOBEC mutagenesis and kataegis were associated with six signatures, highlighting these as a common feature of CIN[39] (Extended Data Fig. 8b and Supplementary Figs. 18 and 21).

Drug response prediction and drug target identification

The putative signature aetiologies implicated canonical cancer pathways as some of the major drivers of CIN. Many of these pathways have been the focus of targeted therapy development. Therefore, given that our signatures can be readily measured in tumours from patients, we explored their utility for therapy response prediction and drug target identification. We integrated data from 297 cancer cell lines, including copy number profiling, genome-wide clustered regularly interspaced short palindromic repeat (CRISPR–Cas9) knockout screens, genome-wide RNA interference (RNAi) screens and the profiling relative inhibition simultaneously in mixtures (PRISM) drug repurposing screen (Supplementary Methods). We assessed correlations between signature activities, gene essentiality and sensitivity to drug perturbation of the gene (Fig. 3a).

We identified 40 genes where copy number signature activity was significantly correlated with both genetic and drug perturbation of the target (Fig. 3b and Supplementary Table 56). Among these, several revealed promising new therapeutic avenues for targeting CIN. CX4 (associated with PI3K–AKT activation) was correlated with response to inhibition of *CCND1* via arcyriaflavin-A, which may indicate a therapeutic strategy for reversing tolerance to WGD[40]. CX5, a signature of IHR, predicted response to olaparib via inhibition of *PARP1*. Given that this signature was also correlated with RNAi knockdown of *PARP1*, this may represent a biomarker that is specific to the inhibition of regular protein function rather than PARP trapping[41]. CX9 (associated with replication stress) was correlated with response to multiple kinase inhibitors targeting genes involved in major mitogenic pathways (*EGFR*, *JAK1*, *MET*, *PRKCA* and *PIK3CA*), suggesting that a multikinase inhibitor approach may be suitable for targeting replication stress. Correlation of CX13 (also associated with replication stress) with response to inhibition of *CDK4* may potentially represent a biomarker-led approach for improving CDK4/6 inhibitor-mediated tumour sensitization to immune checkpoint blockade[42].

Copy number signature correlations with gene essentiality scores from both CRISPR and RNAi perturbation screens identified 104 target genes with druggable structures that currently have no targeted therapies in the clinic (Supplementary Table 57). These represent putative synthetic lethal drug targets, 49 of which had evidence of being implicated in CIN-related mechanisms (Fig. 3c). A number of these show promising links between the signature aetiology and potential consequence of target inhibition. CX1 activity was correlated with perturbation of *ACTL6A* (involved in the SWI/SNF complex) and *TERF1* (involved in telomere maintenance), both of which are required for faithful chromosome segregation during mitosis[4,43]. The combined dysregulation of mitosis and telomere elongation machinery associated with CX1 suggests that inhibiting either one of these genes might be a promising therapeutic strategy by creating synthetic lethality. Indeed, inhibition of both genes has been previously suggested to induce cell lethality by generating excessive CIN[44]. CX9 was correlated with perturbation of *BUB1B*, a spindle assembly checkpoint gene recently identified as therapeutically relevant in CIN-high cells measured via WGD status[45] and an aneuploidy score[7]. This association with CX9 suggests that the spindle assembly checkpoint may have a crucial role in tolerating mid-level amplifications, and reducing levels of *BUB1B* may induce excessive and catastrophic chromosome missegregation[46]. Finally, CX11, which was strongly associated with *CDK4* amplification, was correlated with inhibition of *GNL2*, which in turn impedes the formation of the cyclin D1–CDK4 complex[47].

Predicting platinum sensitivity

The aetiologies of the three IHR signatures suggested a model of increasing CIN complexity (Fig. 4a and Extended Data Fig. 9). IHR alone gives rise to CX2, a signature of small copy number changes indicative of tandem duplication. IHR plus replication stress leads to CX5, which involves larger CNAs. Finally, IHR plus replication stress, impaired damage sensing and impaired NER gives rise to CX3 with the largest CNAs that are strongly associated with loss of heterozygosity. Our results did not reveal whether the different levels of complexity developed in a stepwise manner or by independent processes.

Disruption of both HR[11] and NER[48] have been shown to confer sensitivity to platinum-based chemotherapy. Given that only CX3 was associated with disruption of NER, we hypothesized that the IHR signatures may demonstrate differing abilities to predict platinum sensitivity. As patients with ovarian cancer are routinely treated with platinum-based chemotherapy, we tested the ability of all three signatures to predict overall survival, and hence platinum sensitivity, using a Cox proportional hazards model (Fig. 4b and Supplementary Fig. 32). CX2 showed no association with platinum sensitivity, CX5 was predictive of resistance and CX3 was predictive of sensitivity.

Given that these IHR signatures were able to dissect platinum response, we further hypothesized that they could be used in combination to provide better predictors of platinum sensitivity. As CX2 was not predictive, we used it as a baseline for capturing non-predictive IHR-related genomic changes, and required that the predictive CX3 activity exceed it to potentially confer sensitivity. This resulted in a simple classification rule: 'if CX3 activity is greater than CX2 activity, then predict sensitivity' (Fig. 4c). This interpretable classifier was able to distinguish significant overall survival separation across cohorts of *BRCA1* germline mutant ovarian cancers, ovarian cancers from the TCGA cohort, an independent validation cohort and an oesophageal cancer cohort (also routinely treated with platinum-based chemotherapy) (Fig. 4d, Extended Data Fig. 10 and Supplementary Figs. 33–36). Other classifiers using all three IHR signatures, including more complex machine learning methods, did not outperform this decision rule (Supplementary Fig. 37). Furthermore, this simple classifier had comparable performance to more complex state-of-the-art HRD predictors, which rely on additional data beyond copy number, applied to cohorts of ovarian, oesophageal and breast cancers (Extended Data Fig. 10c, d). By applying this classifier to the whole TCGA ovarian cohort, we estimate that 27% of ovarian tumours might be platinum sensitive. Applying the classifier pan-cancer, we estimate that 8% of all tumours might be sensitive.

Discussion

Here we present a robust analysis framework for CIN in human cancers built on a pan-cancer analysis across 33 cancer types. This resource advances the field in two ways: it untangles CIN according to characteristic genomic patterns and underlying causes, and defines copy number signatures as new biomarkers to quantitatively measure different types of CIN. Our approach complements previous landscape studies of the genetic consequences of CIN[49], which generally focused on recurrent somatic copy number events at individual loci. By contrast, copy number signatures[8,9] uncover mechanistic biases in the patterns of alterations across all chromosomes.

In its current form, the signature methodology cannot account for selection pressures on CNAs. For single-nucleotide variant signatures, passenger mutations provide strong signals for detection. However, for CNAs, the distinction between driver and passenger mutations is less clear. For example, large homozygous deletions are likely to be subject to strong negative selection, whereas other CNAs can be subject to strong positive selection. This has implications for the ability to detect signatures of CIN. Those processes that generate CNAs under positive selection will be easier to detect than those that generate CNAs under negative selection. Quantitatively, the relationship between signature detection and selection is not yet well understood and will depend on genomic background. For example, negative selection will be weaker in whole-genome duplicated samples (approximately 50% of tumours) and in tumours that have lost their ability to sense DNA damage (for example, via *TP53* mutation).

To maximize sample size, we used SNP 6.0 technology data from the TCGA collection. This technology is well established for copy number analysis, but has lower resolution than WGS. As further WGS data become available, there will be an opportunity to refine our signatures and increase their resolution. In their current form, we have demonstrated that the signatures are widely applicable across technologies, including inexpensive assays such as shallow WGS that can be easily applied in a clinical setting to formalin-fixed tumour material[50]. However, it is important to note that the bulk-DNA samples that we analysed do not show dynamics of CIN, and future work is needed to extend our approach to multiple samples or single cells from the same patient to show how patterns of CIN change over time. Further work is also required to quantify copy number signature activity at specific genomic loci, as our method currently only supports signature quantification at a whole-genome level.

The 17 copy number signatures and their putative aetiologies provide a valuable resource for furthering our understanding of CIN. For example, CX1 represents the most prevalent type of CIN across tumours: chromosome missegregation. Aetiology analysis of CX1 pointed at multiple different mitotic defects giving rise to this signature. This suggests that, despite diversity in the potential causes of mitotic defects, these all result in the same change in genome structure[1]. These missegregation events typically result in large copy number changes, potentially disrupting the function of many genes; however, our signature analysis reveals that these changes only represent, on average, 4% of the total number of copy number changes observed in a tumour (Supplementary Fig. 38). By contrast, CX2 accounts for 23% of the copy number changes observed in a tumour. This highlights the power of our compendium of signatures to quantify and disentangle the causes and functional effect that different types of CIN have on tumour genomes. Our results also highlight the potential of our signatures to improve the treatment of patients with extreme CIN tumours. Platinum-based chemotherapy is currently considered a broad-spectrum cytotoxic chemotherapy and is routinely used to treat cancers with extreme CIN. However, here we showed that platinum response can be robustly dissected using different signatures of IHR. By developing the IHR signatures into a companion diagnostic assay, platinum-based therapies could potentially be administered in a more targeted manner, allowing resistant patients to avoid their toxic side effects, and healthcare systems to reduce the cost burden of ineffectual treatment. Similarly for other signatures, our analysis of drug response across cell lines reinforces their potential to be developed into companion diagnostics for improved patient stratification during clinical trials.

The signature compendium presented here is an important resource to guide future studies into a deeper understanding of the origins and diversity of CIN and how to therapeutically target different types of CIN.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04789-9.

1. Bakhoum, S. F. & Cantley, L. C. The multifaceted role of chromosomal instability in cancer and its microenvironment. *Cell* **174**, 1347–1360 (2018).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
3. Tijhuis, A. E., Johnson, S. C. & McClelland, S. E. The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. *Mol. Cytogenet.* **12**, 17 (2019).
4. Chakravarti, D., LaBella, K. A. & DePinho, R. A. Telomeres: history, health, and hallmarks of aging. *Cell* **184**, 306–322 (2021).
5. Bakhoum, S. F. et al. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).
6. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
7. Cohen-Sharir, Y. et al. Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature* **590**, 486–491 (2021).
8. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
9. Steele, C. D. et al. Undifferentiated sarcomas develop through distinct evolutionary pathways. *Cancer Cell* **35**, 441–456.e8 (2019).
10. Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).
11. Stok, C., Kok, Y. P., van den Tempel, N. & van Vugt, M. A. T. M. Shaping the BRCAness mutational landscape by alternative double-strand break repair, replication stress and mitotic aberrancies. *Nucleic Acids Res.* **49**, 4239–4257 (2021).
12. Takemon, Y. *et al*. Multi-omic analyses reveal a role for mammalian CIC in cell cycle regulation and mitotic fidelity Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/533323v2 (2019).
13. Hell, M. P., Duda, M., Weber, T. C., Moch, H. & Krek, W. Tumor suppressor VHL functions in the control of mitotic fidelity. *Cancer Res.* **74**, 2422–2431 (2014).
14. Brownlee, P. M., Chambers, A. L., Cloney, R., Bianchi, A. & Downs, J. A. BAF180 promotes cohesion and prevents genome instability and aneuploidy. *Cell Rep.* **6**, 973–981 (2014).
15. Silverman, J. S., Skaar, J. R. & Pagano, M. SCF ubiquitin ligases in the maintenance of genome stability. *Trends Biochem. Sci.* **37**, 66–73 (2012).
16. Godinho, S. A. & Pellman, D. Causes and consequences of centrosome abnormalities in cancer. *Phil. Trans. R. Soc. B* **369**, 20130467 (2014).
17. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210.e5 (2018).
18. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **107**, 1776–1782 (2012).
19. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
20. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
21. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
22. Ogden, A., Rida, P. C. G. & Aneja, R. Prognostic value of CA20, a score based on centrosome amplification-associated genes, in breast tumors. *Sci Rep.* **7**, 262 (2017).
23. Piazza, A. & Heyer, W.-D. Homologous recombination and the formation of complex genomic rearrangements. *Trends Cell Biol.* **29**, 135–149 (2019).
24. Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Front. Genet.* **5**, 175 (2014).
25. Knijnenburg, T. A. et al. Genomic and molecular landscape of dna damage repair deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239–254.e6 (2018).
26. Saavedra, H. I., Fukasawa, K., Conn, C. W. & Stambrook, P. J. MAPK mediates RAS-induced chromosome instability. *J. Biol. Chem.* **274**, 38083–38090 (1999).
27. Perl, A. L. et al. Protein phosphatase 2A controls ongoing DNA replication by binding to and regulating cell division cycle 45 (CDC45). *J. Biol. Chem.* **294**, 17043–17059 (2019).
28. Chen, L. et al. The augmented R-loop is a unifying mechanism for myelodysplastic syndromes induced by high-risk splicing factor mutations. *Mol. Cell* **69**, 412–425.e6 (2018).
29. Li, Q. et al. ERCC2 helicase domain mutations confer nucleotide excision repair deficiency and drive cisplatin sensitivity in muscle-invasive bladder cancer. *Clin. Cancer Res.* **25**, 977–988 (2019).
30. Menon, V. & Povirk, L. Involvement of p53 in the repair of DNA double strand breaks: multifaceted roles of p53 in homologous recombination repair (HRR) and non-homologous end joining (NHEJ). *Subcell. Biochem.* **85**, 321–336 (2014).
31. Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239–2254.e39 (2021).
32. Dewhurst, S. M. et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
33. Davoli, T. & de Lange, T. The causes and consequences of polyploidy in normal development and cancer. *Annu. Rev. Cell Dev. Biol.* **27**, 585–610 (2011).
34. Berenjeno, I. M. et al. Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. *Nat. Commun.* **8**, 1773 (2017).
35. Darp, R., Vittoria, M. A., Ganem, N. J. & Ceol, C. J. Oncogenic BRAF induces whole-genome doubling through suppression of cytokinesis. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.08.439023 (2021).
36. Zhang, Q. et al. FBXW7 facilitates nonhomologous end-joining via K63-linked polyubiquitylation of XRCC4. *Mol. Cell* **61**, 419–433 (2016).
37. Citri, A., Skaria, K. B. & Yarden, Y. The deaf and the dumb: the biology of ErbB-2 and ErbB-3. *Exp. Cell Res.* **284**, 54–65 (2003).

38. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
39. Venkatesan, S. et al. Induction of APOBEC3 exacerbates DNA replication stress and chromosomal instability in early breast and lung cancer evolution. *Cancer Discov.* **11**, 2456–2473 (2021).
40. Crockford, A. et al. Cyclin D mediates tolerance of genome-doubling in cancers with functional p53. *Ann. Oncol.* **28**, 149–156 (2017).
41. Ray Chaudhuri, A. & Nussenzweig, A. The multifaceted roles of PARP1 in DNA repair and chromatin remodelling. *Nat. Rev. Mol. Cell Biol.* **18**, 610–621 (2017).
42. Goel, S. et al. CDK4/6 inhibition triggers anti-tumour immunity. *Nature* **548**, 471–475 (2017).
43. Brownlee, P. M., Meisenberg, C. & Downs, J. A. The SWI/SNF chromatin remodelling complex: Its role in maintaining genome stability and preventing tumourigenesis. *DNA Repair* **32**, 127–133 (2015).
44. Kops, G. J. P., Foltz, D. R. & Cleveland, D. W. Lethality to human cancer cells through massive chromosome loss by inhibition of the mitotic checkpoint. *Proc. Natl Acad. Sci. USA* **101**, 8699–8704 (2004).
45. Quinton, R. J. et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **590**, 492–497 (2021).
46. Janssen, A., Kops, G. J. P. L. & Medema, R. H. Elevating the frequency of chromosome mis-segregation as a strategy to kill tumor cells. *Proc. Natl Acad. Sci. USA* **106**, 19108–19113 (2009).
47. Datta, D. et al. Nucleolar GTP-binding protein-1 (NGP-1) promotes G1 to S phase transition by activating cyclin-dependent kinase inhibitor p21 Cip1/Waf1. *J. Biol. Chem.* **290**, 21536–21552 (2015).
48. Martin, L. P., Hamilton, T. C. & Schilder, R. J. Platinum resistance: the role of DNA repair pathways. *Clin. Cancer Res.* **14**, 1291–1295 (2008).
49. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
50. 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**, 2022–2032 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reporting summary

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

Data availability

All data used in this study were obtained from publicly available sources and are described in detail in Supplementary Table 1, section 'Data and Code' in the Supplementary Methods. Some raw data have restricted access (TCGA dbGaP accession number: phs000178.v11.p8; ICGC EGA accession number: EGAS00001001692). Access can be obtained by applying to the relevant Data Access Committees (TCGA or ICGC). The authors declare that all other data supporting the findings of this study, including the source data for all figures, are publicly available without restrictions and are also available in the Supplementary Information and the Github repositories. All data supporting the analysis of our copy number signatures are navigable via our web portal (https://markowetz.cruk.cam.ac.uk/cincompendium/).

Code availability

The code is publicly accessible via our hub repository (https://github.com/markowetzlab/Drews2022_CIN_Compendium), which describes how the CIN signatures were derived and how to reproduce the figures and tables in this publication. The repository also contains the publicly accessible data and intermediary results used and produced in this study. The hub repository links to other repositories containing the code for specialized tasks.

Acknowledgements We thank M. Eldridge for setting up the online resource and A. Berman for bug fixes. R.M.D., P.S.S., D.-L.C., J.D.B. and F.M. are funded by Cancer Research UK (core grants C14303/A17197, A22905 and A19274) and the Cambridge Cancer Centre (grant C9685/A25117). G.M. and B.H. are hosted by the Centro Nacional de Investigaciones Oncológicas (CNIO), which is supported by the Instituto de Salud Carlos III and recognized as a 'Severo Ochoa' Centre of Excellence (ref. CEX2019-000891-S) by the Spanish Ministry of Science and Innovation (MCIN/AEI/ 10.13039/501100011033). G.M. and B.H. were also supported by a Spanish Ministry of Science and Innovation grant PID2019-111356RA-I00 (MCIN/AEI/ 10.13039/501100011033). M.T. was supported as a postdoctoral researcher of the F.R.S.-FNRS. L.M.G. was supported by the Wellcome Trust PhD programme in Mathematical Genomics and Medicine (grant number RG92770). M.S. was supported by the Horizon 2020 (H2020) Integrated Training Network CONTRA (grant 766030-CONTRA-H2020-MSCA-ITN-2017). This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute.

Author contributions G.M. and F.M. contributed equally to this work. R.M.D., G.M. and F.M. conceived and designed the study. R.M.D., B.H., D.-L.C., M.S. and G.M. developed the methodology of the study. R.M.D., B.H., M.T., K.H., T.L., P.S.S., L.M.G., L.L., M.S. and G.M. developed the software for the study. R.M.D., M.T., K.H., T.L., P.S.S., D.-L.C. and G.M. contributed to the validation of the method and results. R.M.D., B.H., L.M.G., D.-L.C., L.L. and G.M. contributed to the formal analysis presented in this study. R.M.D., B.H., K.H., T.L., P.S.S., L.M.G. and P.V.L. provided access to data and contributed to gathering, processing and curating data. R.M.D., J.D.B., P.V.L., G.M. and F.M. wrote the original draft. R.M.D., B.H., G.M. and F.M. produced and contributed to the visualizations of the study. R.M.D., G.M. and F.M. supervised the project. All authors had access to all of the data in the study. All authors contributed to the review and the editing of the manuscript. All authors approved the manuscript before the initial submission and all other resubmissions.

Competing interests J.D.B., G.M., F.M. are co-founders, directors and share holders of Tailor Bio Ltd. Cancer Research UK, the University of Cambridge and the Spanish National Cancer Research Centre (CNIO) are in the process of applying for a patent application GB2114203.9 covering the copy number signature methodology and drug target associations that lists R.M.D., B.H., G.M. and F.M. as inventors. G.M., F.M. and J.D.B are listed on a patent on using copy number signatures to predict response to doxorubicin treatment in ovarian cancer (PCT/EP2021/065058).

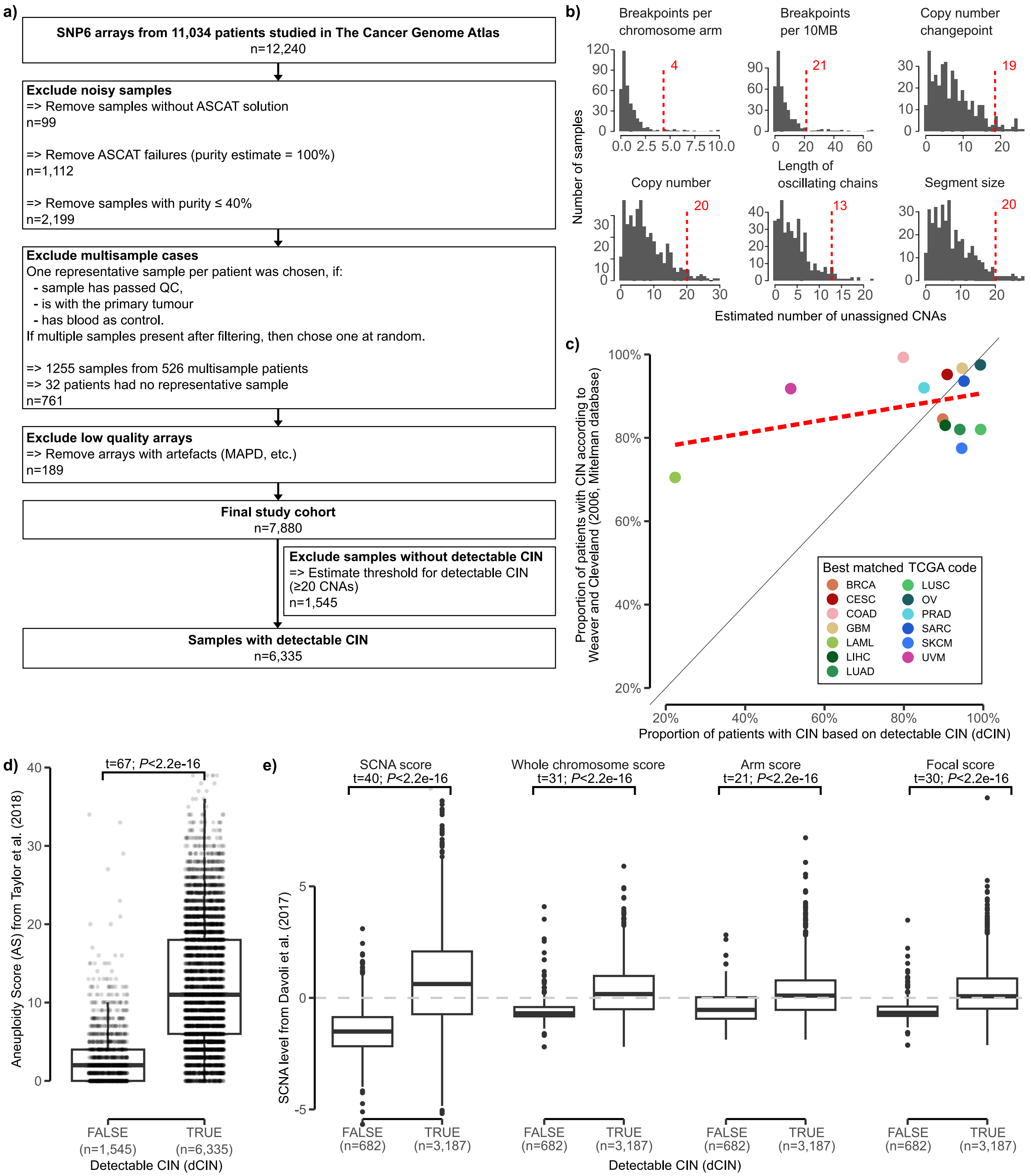
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04789-9.

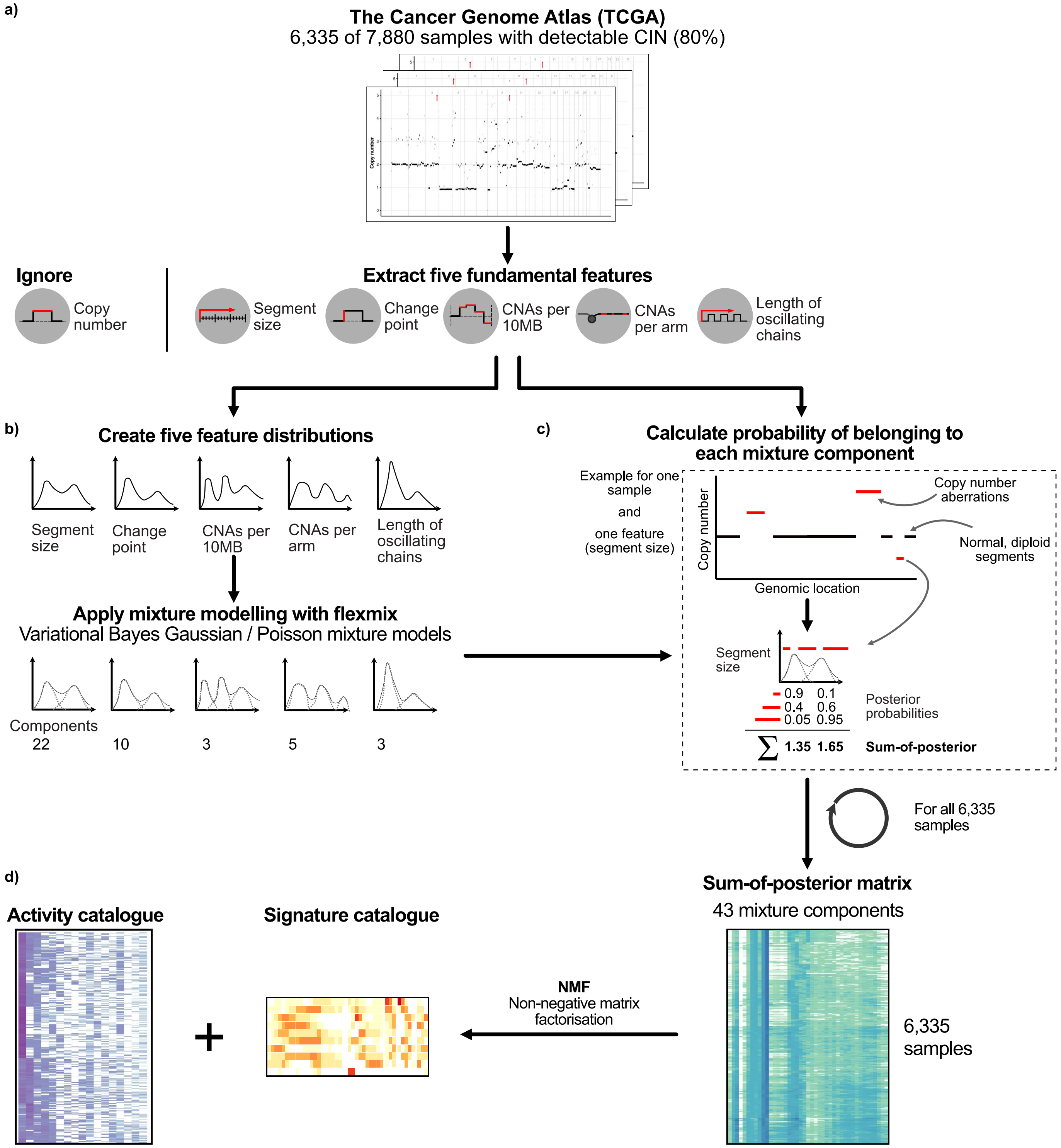
Correspondence and requests for materials should be addressed to Geoff Macintyre or Florian Markowetz.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

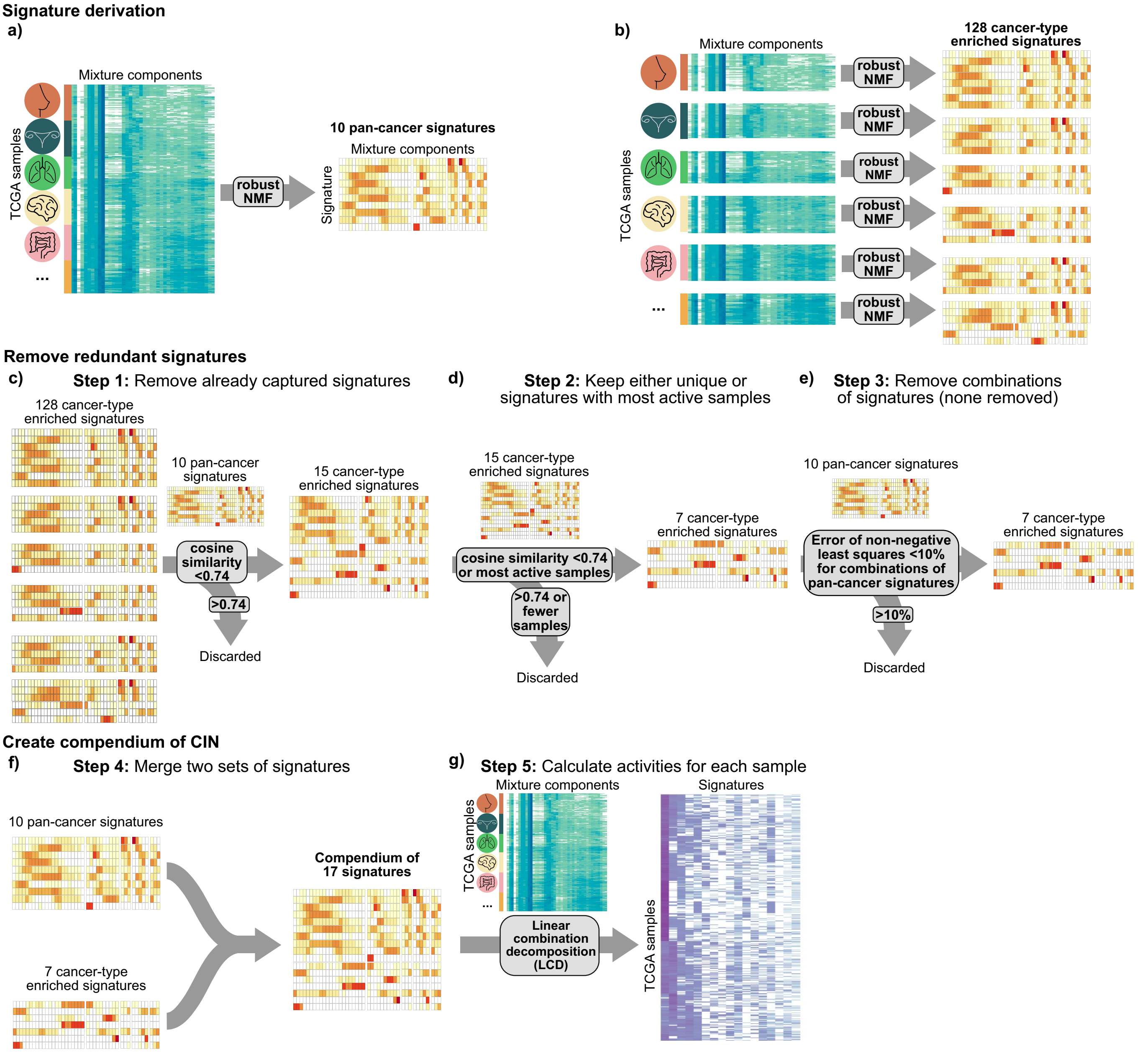
Reprints and permissions information is available at http://www.nature.com/reprints.



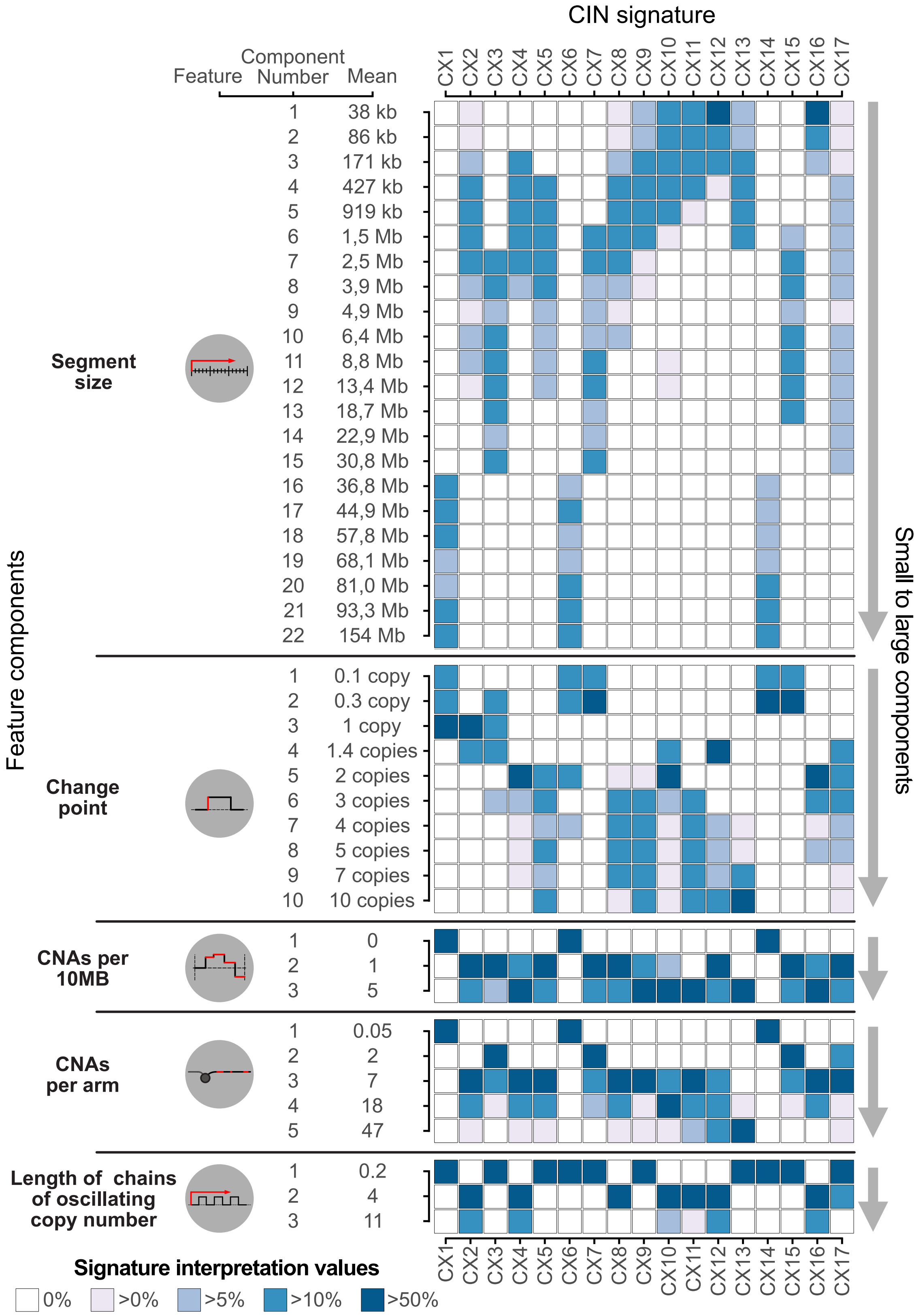
Extended Data Fig. 1 | Workflow of sample filtering and detectable chromosomal instability (dCIN). **a**, REMARK diagram showing flow of samples through the study. **b**, For each copy number feature of the previous ovarian signatures: a histogram of number of events per sample that could not be assigned to an ovarian copy number signature on the TCGA ovarian cohort. Red dotted line indicates the quantile 0.95. **c**, Scatterplot of cancer types comparing our estimate of detectable CIN (Supplementary Methods) to estimates reported in the Mitelman database. **d**,**e**, Boxplots comparing our estimate of detectable CIN with aneuploidy score and four CNA-specific metrics. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points. Results of two-sided Welch's t-test shown on top of the boxplots.



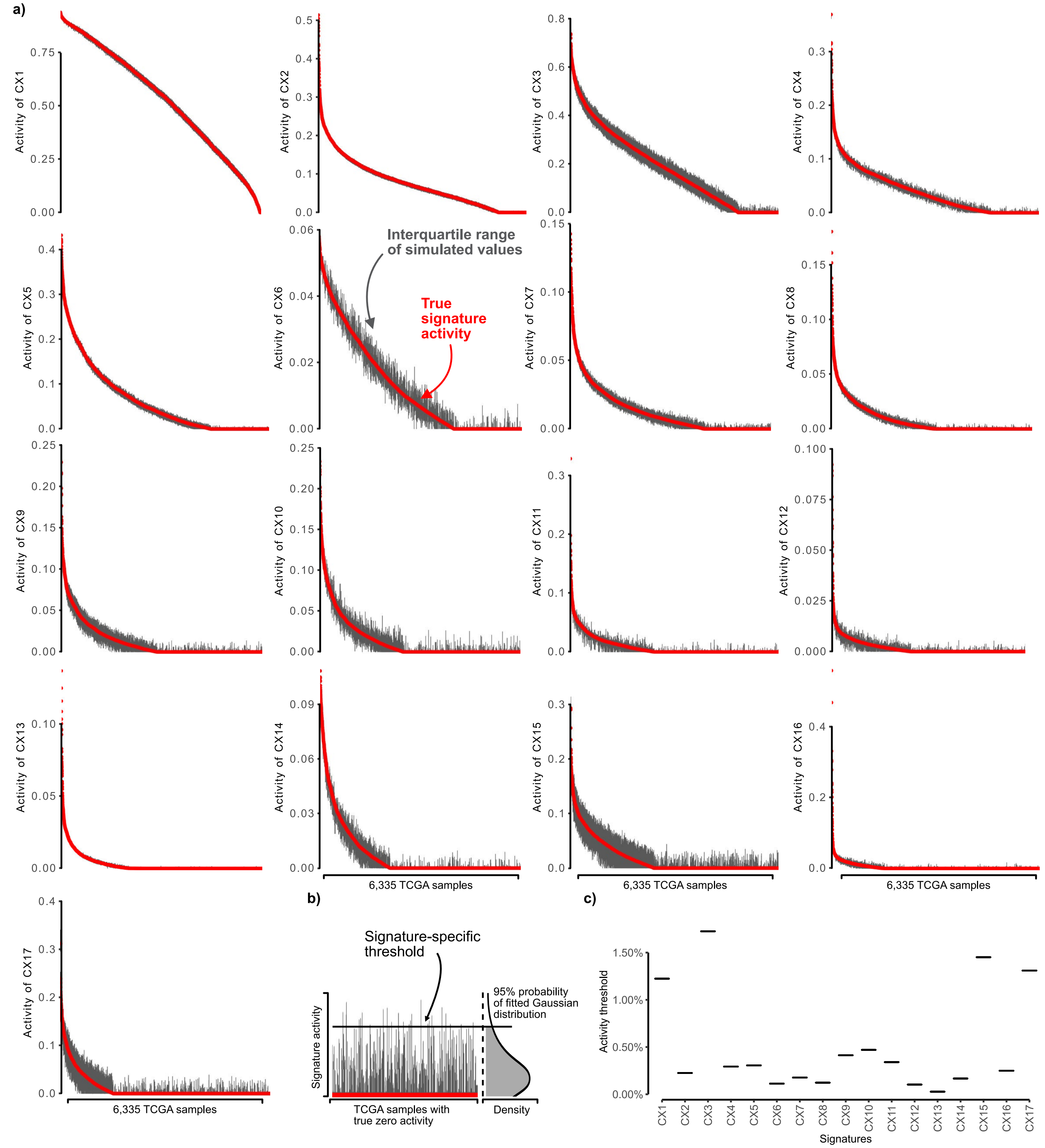
Extended Data Fig. 2 | Overview of copy number features and signature identification. a, A schematic showing the 5 fundamental copy number features that were computed using 6,335 samples with detectable CIN (dCIN). Note, a feature capturing absolute copy number is not included in our method. **b**, A schematic showing how mixture modelling is used to split the genome-wide feature distributions into smaller components by either Variational Bayes Gaussian mixture models or Finite Poisson mixture models. The actual number of resulting components is listed below each feature distribution. These components represent basic building blocks of each feature distribution. **c**, An example of how the probability of a CNA belonging to a mixture component (posterior probability) is calculated and how these are summed. **d**, (Right) The resulting 43-dimensional feature vectors for each sample, after all posterior probabilities are summed for each component. (Left) A schematic of how the sum-of-posterior matrix for all 6,335 samples was split in two matrices by a Bayesian implementation of the non-negative matrix factorisation (NMF), resulting in a signature catalogue and an activity catalogue.


Signature derivation
a)
Mixture components
TCGA samples
...
robust NMF
10 pan-cancer signatures
Mixture components
Signature
b)
Mixture components
TCGA samples
...
robust NMF
128 cancer-type enriched signatures
Remove redundant signatures
c)
Step 1: Remove already captured signatures
128 cancer-type enriched signatures
10 pan-cancer signatures
cosine similarity <0.74
>0.74
Discarded
15 cancer-type enriched signatures
d)
Step 2: Keep either unique or signatures with most active samples
15 cancer-type enriched signatures
cosine similarity <0.74 or most active samples
>0.74 or fewer samples
Discarded
7 cancer-type enriched signatures
e)
Step 3: Remove combinations of signatures (none removed)
10 pan-cancer signatures
Error of non-negative least squares <10% for combinations of pan-cancer signatures
>10%
Discarded
7 cancer-type enriched signatures
Create compendium of CIN
f)
Step 4: Merge two sets of signatures
10 pan-cancer signatures
7 cancer-type enriched signatures
Compendium of 17 signatures
g)
Step 5: Calculate activities for each sample
Mixture components
TCGA samples
...
Linear combination decomposition (LCD)
Signatures
TCGA samples

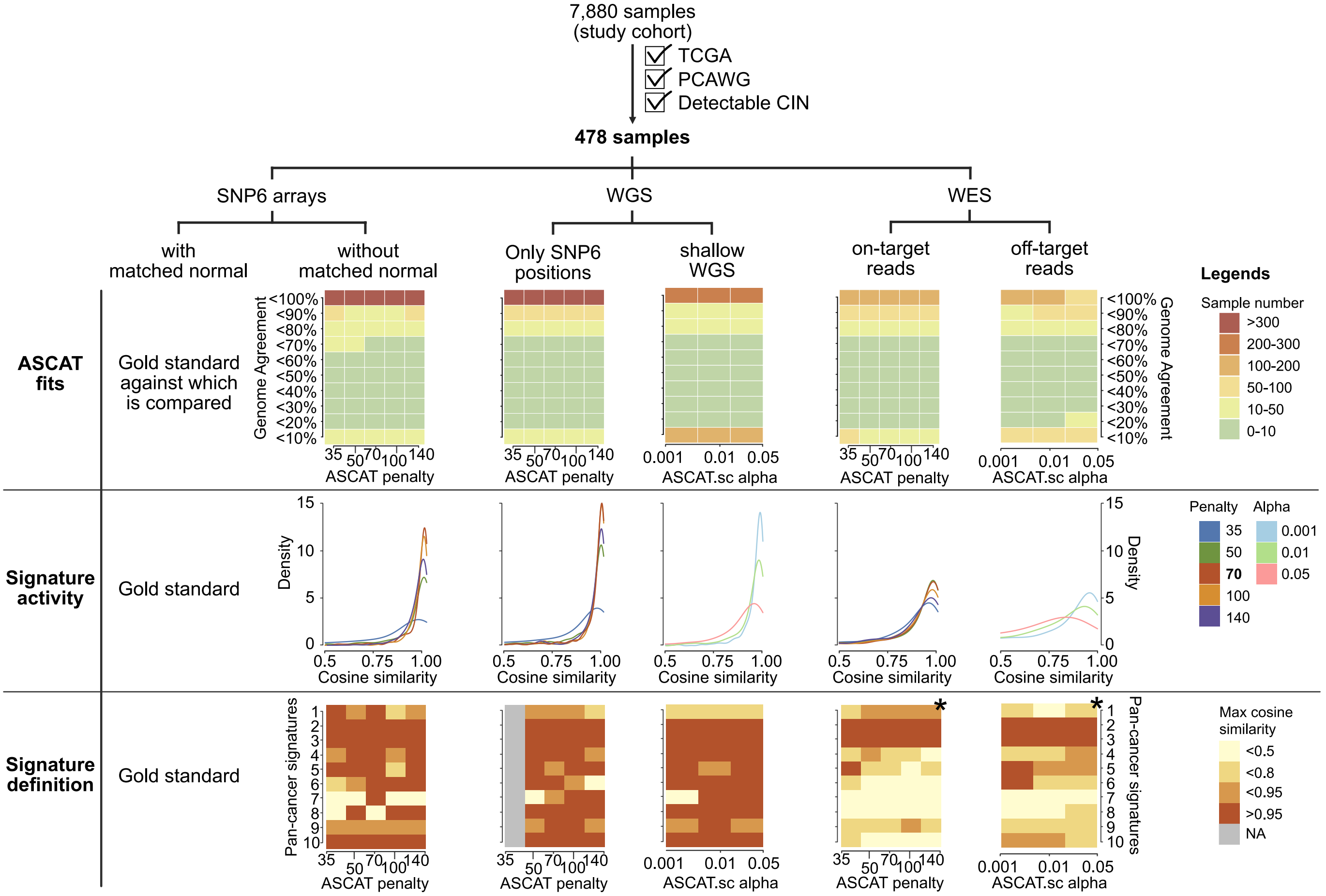
Extended Data Fig. 3 | Schematic of the signature compendium identification. **a**, From the complete input matrix 10 pan-cancer signatures were identified. **b**, For the 20 cancer types with over 100 samples each, 128 cancer-type enriched signatures (CTES) were identified. **c**, All CTES were removed that had a cosine similarity over 0.74 with any pan-cancer signature. **d**, From the groups of CTES that had cosine similarities over 0.74 to each other, the signature with activities in the largest number of samples was taken as a representative signature. **e**, We performed non-negative least squares on each pair of pan-cancer specific signatures to each CTES. For any combination which showed a reconstruction error below 0.1, this CTES was removed. **f**, The sets of 10 pan-cancer and 7 CTES were joined to a compendium of 17 signatures. **g**, Using linear combination decomposition, the signature activities were calculated for the 6,335 TCGA samples.



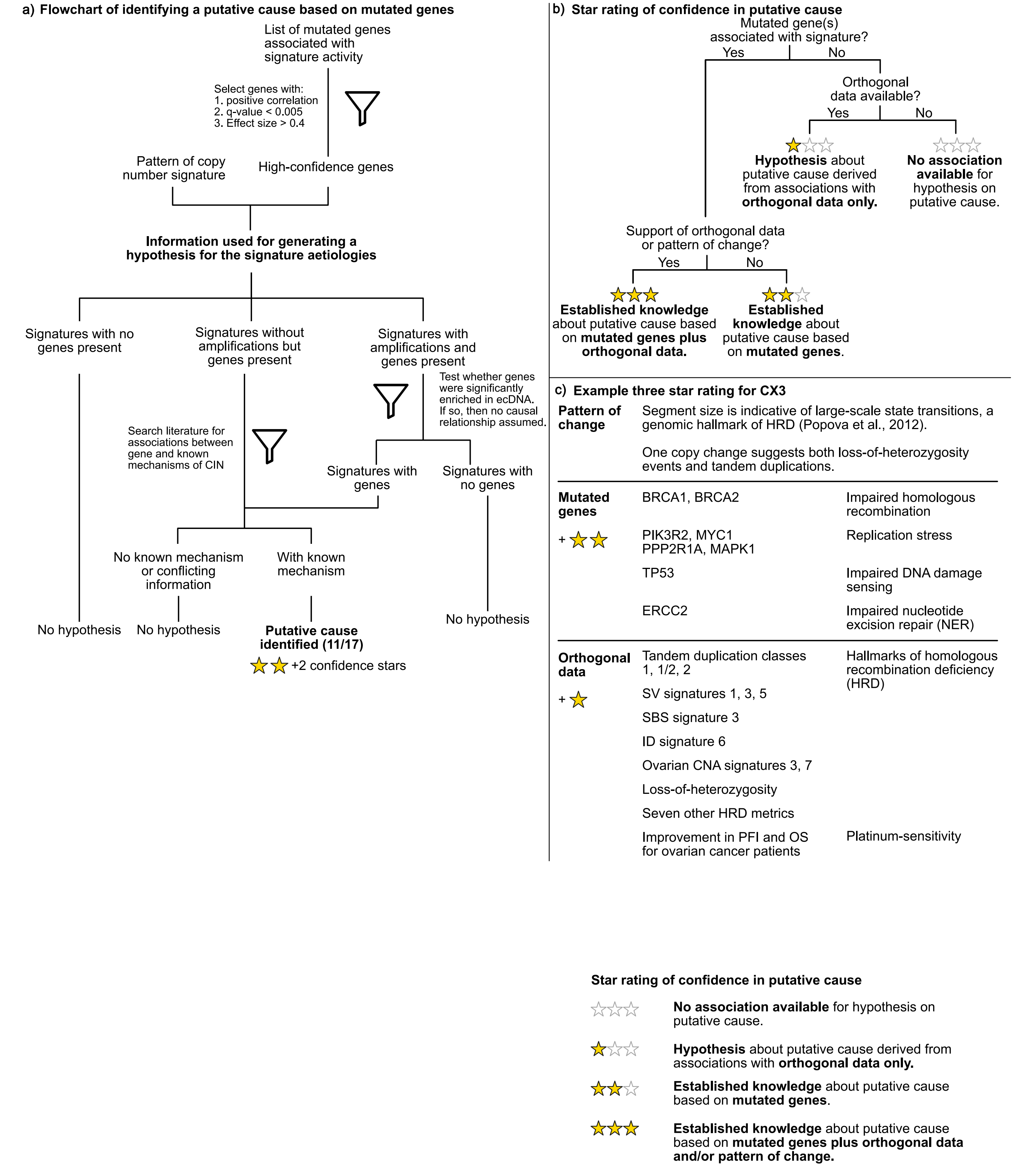
Extended Data Fig. 4 | Signature interpretation matrix. Displayed on the left are the five features, their mixture components and component means. The heatmap on the right shows the signature interpretation values, which combine information from the sum-of-posterior matrix, signature activity matrix and the signature definition matrix (Supplementary Methods). Only components that are positively correlated with signature activity levels are displayed. Interpretation values are normalised per feature and signature.



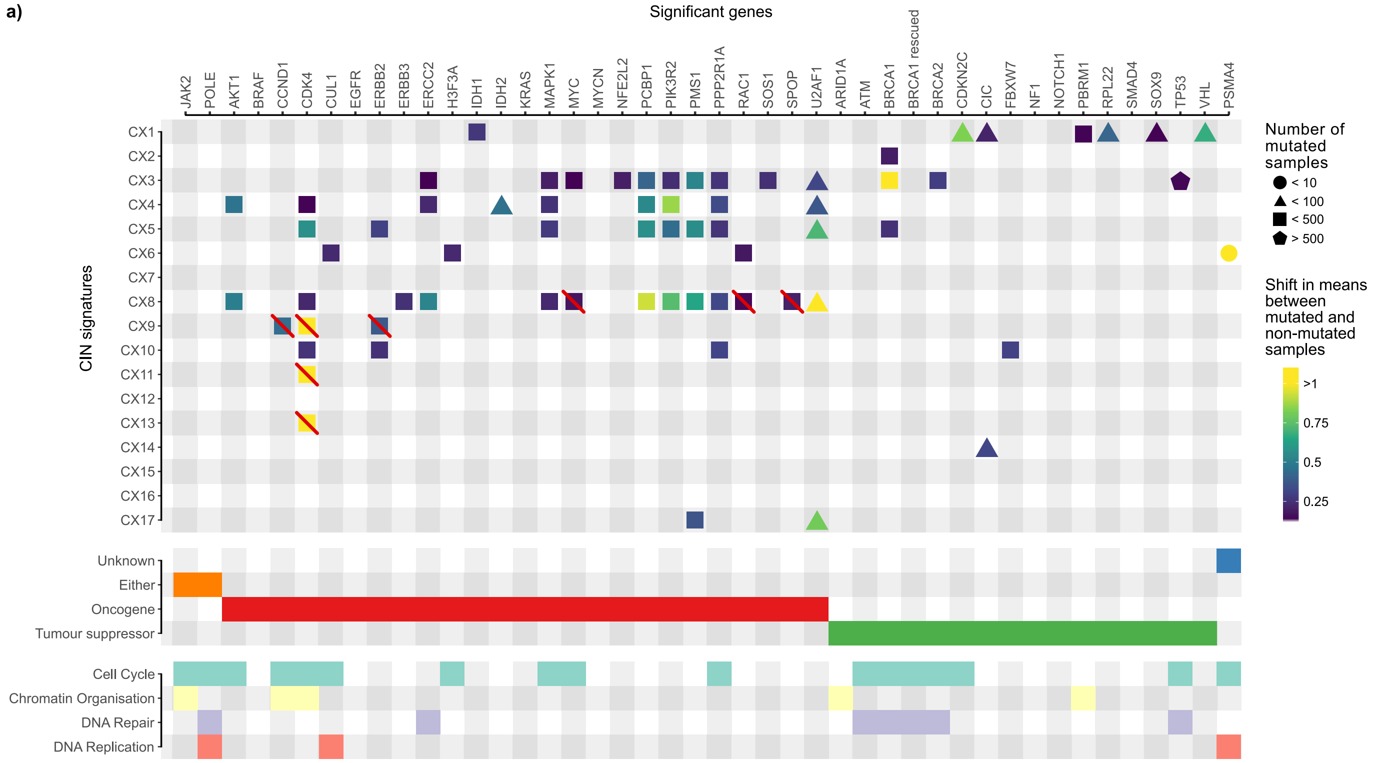
Extended Data Fig. 5 | Monte Carlo simulation results for determining signature-specific noise thresholds. a, Each plot (1 per signature) shows the interquartile range of sample signature activities after the introduction of noise in the copy number features using a Monte Carlo simulation. Samples are ordered by their observed signature activity (red line). **b**, Schematic showing how we fitted a Gaussian distribution to the simulated values of all samples with an observed signature activity of 0 (red line). The horizontal black line represents the quantile 0.95 of the fitted Gaussian and forms the basis of our signature specific noise threshold, where values below this line are not distinguishable from 0. **c**, Plot of the signature-specific thresholds for the 17 copy number signatures.



Extended Data Fig. 6 | Signature stability across different copy number profiling technologies. Across the same set of 478 tumours, we compared the SNP6-array based copy number profiles and signatures to copy number profiles and signatures derived using different copy number profiling technologies. The columns contain results for the different technologies and the rows contain results for comparison between copy number profiles (top), signature activities (middle) and signature definitions (bottom, limited to pan-cancer signatures). For each comparison we show results for a range of penalties for ASCAT's piecewise constant fitting or ASCAT.sc's circular binary segmentation. (*): For settings marked with a star it was not possible to derive solutions for K = 10, instead the optimal number of K was chosen (lower than K = 10).


a) Flowchart of identifying a putative cause based on mutated genes
List of mutated genes associated with signature activity
Select genes with:
1. positive correlation
2. q-value < 0.005
3. Effect size > 0.4
Pattern of copy number signature
High-confidence genes
Information used for generating a hypothesis for the signature aetiologies
Signatures with no genes present
Signatures without amplifications but genes present
Signatures with amplifications and genes present
Test whether genes were significantly enriched in ecDNA. If so, then no causal relationship assumed.
Search literature for associations between gene and known mechanisms of CIN
Signatures with genes
Signatures with no genes
No known mechanism or conflicting information
With known mechanism
No hypothesis
No hypothesis
Putative cause identified (11/17)
★★ +2 confidence stars
No hypothesis
b) Star rating of confidence in putative cause
Mutated gene(s) associated with signature?
Yes
No
Orthogonal data available?
Yes
No
★☆☆
Hypothesis about putative cause derived from associations with orthogonal data only.
☆☆☆
No association available for hypothesis on putative cause.
Support of orthogonal data or pattern of change?
Yes
No
★★★
Established knowledge about putative cause based on mutated genes plus orthogonal data.
★★☆
Established knowledge about putative cause based on mutated genes.
c) Example three star rating for CX3
Pattern of change
Segment size is indicative of large-scale state transitions, a genomic hallmark of HRD (Popova et al., 2012).
One copy change suggests both loss-of-heterozygosity events and tandem duplications.
Mutated genes
+ ★★
BRCA1, BRCA2
Impaired homologous recombination
PIK3R2, MYC1 PPP2R1A, MAPK1
Replication stress
TP53
Impaired DNA damage sensing
ERCC2
Impaired nucleotide excision repair (NER)
Orthogonal data
+ ★
Tandem duplication classes 1, 1/2, 2
Hallmarks of homologous recombination deficiency (HRD)
SV signatures 1, 3, 5
SBS signature 3
ID signature 6
Ovarian CNA signatures 3, 7
Loss-of-heterozygosity
Seven other HRD metrics
Improvement in PFI and OS for ovarian cancer patients
Platinum-sensitivity
Star rating of confidence in putative cause
☆☆☆
No association available for hypothesis on putative cause.
★☆☆
Hypothesis about putative cause derived from associations with orthogonal data only.
★★☆
Established knowledge about putative cause based on mutated genes.
★★★
Established knowledge about putative cause based on mutated genes plus orthogonal data and/or pattern of change.

Extended Data Fig. 7 | Workflow for determining signature aetiology and confidence rating. a, Flowchart showing how an association between a mutated gene and signature activity was used to derive a hypothesis for a putative aetiology. **b**, Flowchart representing the decision making process leading to the assignment of a 3-star rating confidence score. **c**, Example of the star rating process for CX3.



b)

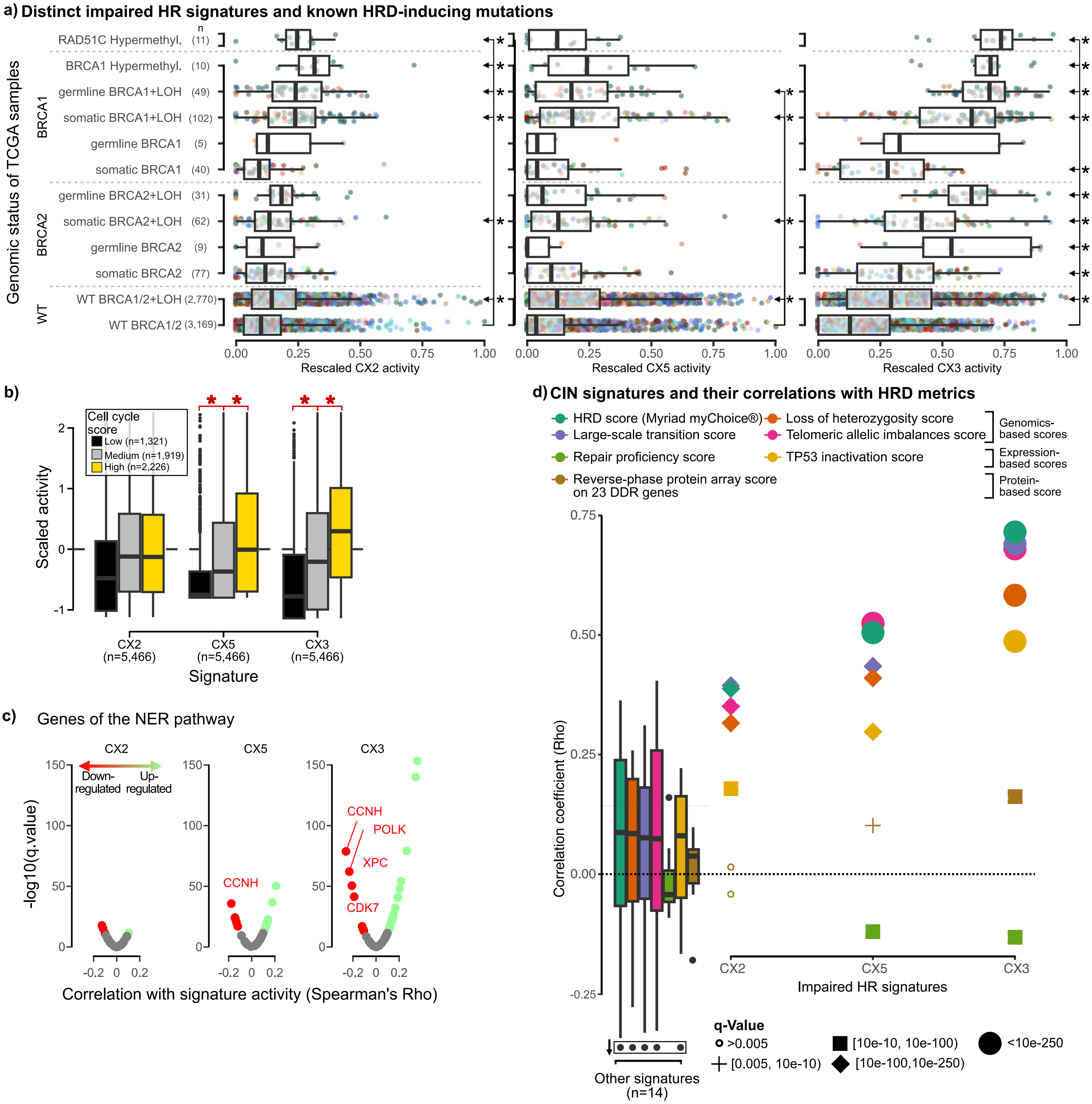
CIN signatures

Biological covariates

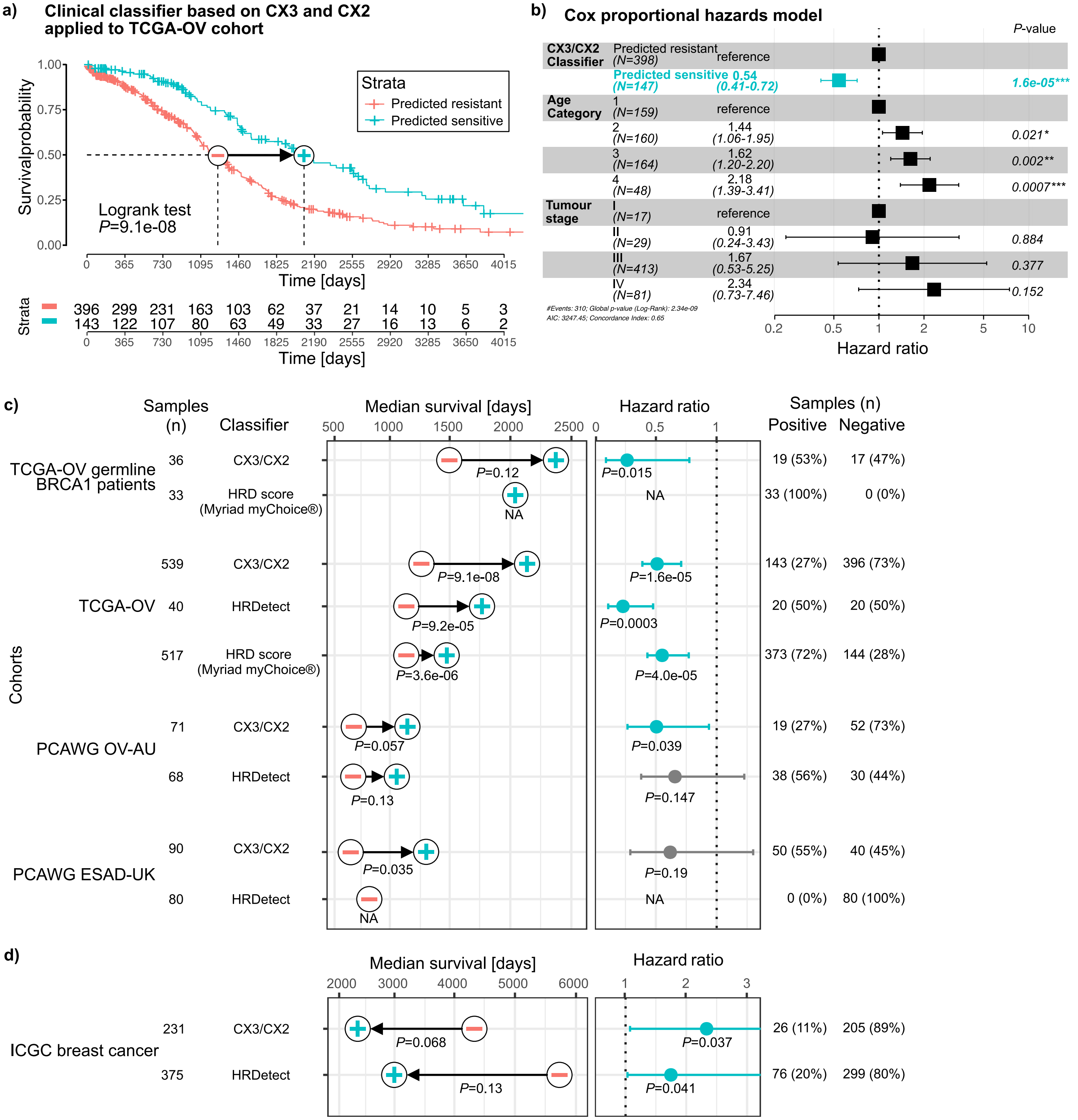
	CX1	CX2	CX3	CX4	CX5	CX6	CX7	CX8	CX9	CX10	CX11	CX12	CX13	CX14	CX15	CX16	CX17
Whole-chromosome CNAs	↑			↑		↑								↑			
Cell cycle score			↑	↑	↑		↑	↑	↑		↑		↑				
Chromothripsis		↑		↑	↑			↑	↑	↑			↑				
Circular ecDNA													↑				
Kataegis		↑	↑	↑	↑			↑	↑				↑				
Loss of heterozygosity		↑	↑		↑			↑									
SBS signature	1 5	2 3 13	2 3 13	2 13	2 3 13	5		2 13	2 13	2			2 13			3	
ID signature	1	2 6 8 9	2 6 8	2	2 6 9			6 8	1 2	2	1 2		2			6	
Rearrangement signature	2	4 6	1 3 5		1 3 4 5 6		3 5	1 6	4 6				4 6	5			
Tandem duplication (TD)		↑	↑		↑			↑		↑							
Dominant TD class		2 3	1		2 3			2		1/2 2						1 2/3	
Whole-genome duplication	–	–	+	+	+	+	–	+	+				+		–		+

Legend ↑ / + Positive significant correlation – Negative significant correlation

Extended Data Fig. 8 | Summary of associations between signatures and other covariates. a, Main panel shows significant associations between copy number signatures and mutated genes. Gene annotations summarised in the panels below. Boxes with a red line indicate significant associations that were not considered when determining signature aetiologies as the significant enrichment was via amplification of the gene, which also resided in an ecDNA amplicon, which could be a consequence of the signature rather than a cause, potentially causing a spurious correlation with amplification signatures (CX8, CX9, CX11, CX13). **b**, Each row shows highly significant associations between signatures and different covariates. Unless otherwise specified, only positive correlations are shown.



Extended Data Fig. 9 | Impaired homologous recombination signatures and their associations. a, Boxplots summarise signature activities of different patient groups (rows) defined by their driver gene mutation status. Ovarian samples are coloured in dark green and breast in orange. Boxes represent the interquartile range (IQR) with the median as a bolded line. The whiskers extend to the largest/smallest value no further than 1.5 * IQR from the hinge. Outliers beyond the end of the whiskers are marked individually as points. Significance tested with two-sided Welch's t-test between WT *BRCA1/2* and each of the categories and corrected for multiple testing by using Benjamini-Hochberg method. Statistically significant comparisons are shown to the right of the boxplots with stars denoting significance ($q < 0.05$) and arrows denoting the two groups used for the statistical test. (*BRCA1/2* = *BRCA1* and *BRCA2*, WT = wild type; LOH = loss of heterozygosity). **b**, Boxplots (with same characteristics as in **a**) summarise the scaled signature activities of 5,466 TCGA samples split by low, medium and high cell cycle scores. The brackets and stars ($q < 0.05$) show where there was a significant increase from low to medium to high cell cycle groups tested with a Welch's t-test and corrected for multiple testing with Benjamini-Hochberg method. **c**, Volcano plots showing the results of a correlation between signature activity and expression of genes involved in nucleotide excision repair (NER). Each dot represents a gene, coloured dots show significant correlations. **d**, Spearman correlation coefficient (y-axis) of correlation between signature activities and seven common metrics of HRD (listed at top). Individual coefficients are displayed for impaired homologous recombination (IHR) signatures and the distribution of coefficients from remaining signatures are represented by boxplots (with same characteristics as in **a**).



Extended Data Fig. 10 | Performance of classifiers for predicting platinum sensitivity. **a**, Kaplan-Meier estimator showing the overall survival probabilities of TCGA ovarian cancer patients split into two groups using our CX3/CX2 classifier. **b**, Hazard ratios and their 95% confidence interval obtained from a Cox proportional hazards model trained on our CX3/CX2 classification predicting overall survival of TCGA ovarian cancer patients. The model also corrected for age and cancer stage of the patients. P-value represents the significance of a Wald test. **c**,**d**, Median survival and hazard ratios generated for five cancer cohorts from the TCGA, PCAWG and ICGC projects using predictions from three classifiers (our CX3/CX2 classifier, HRDetect and Myriad myChoice based on the HRD score). Improvements in median survival tested by log-rank test (Kaplan-Meier survival analysis), with the minus symbol representing the predicted resistant group and the plus symbol the predicted sensitive group. Hazard ratios, their 95% confidence interval, and Wald test significance of the predicted sensitive group compared to the predicted resistant group are obtained from Cox proportional hazards models correcting for stage and age of patients, except for HRDetect where tumour stage was omitted as the models did not converge if included. The number and proportion of patients predicted to be sensitive (with HRD) and resistant (without HRD) by each classifier are listed on the right.

nature research

Corresponding author(s): Ruben Matthias Drews, Geoff Macintyre, Florian Markowetz

Last updated by author(s): Apr 8, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

All software is open to the members of the public and accessible via github. The following links provide unrestricted access:

Main repositories
- Drews2022 CIN Compendium is the main github repository containing code to reproduce the figures, most results and most supplementary tables: https://github.com/markowetzlab/Drews2022_CIN_Compendium
- CINSignatureQuantification covers the R package needed to quantify signature activities from copy number profiles. Also includes clinical classifier: https://github.com/markowetzlab/CINSignatureQuantification
- CINSignatureDiscovery covers 1) how the signature compendium was obtained and 2) details how to identify de-novo signatures from copy number profiles: https://github.com/markowetzlab/CINSignatureDiscovery

Support repositories
- ASCAT is the tool used to generate absolute copy number profiles from Affymetrix SNP6 arrays. The final copy number profiles can be downloaded from the github repository: https://github.com/VanLoo-lab/ascat
- CIN Genome Simulation covers the simulation we used to validate the multi-feature approach: https://github.com/markowetzlab/CINGenomeSimulation
- Drug target discovery and biomarker identification supports the results shown in Figure 3: https://github.com/macintyrelab/CNSigs_DrugCorrelations
- Signature Interpretation Matrix details how the Interpretation matrix was obtained: https://github.com/macintyrelab/SignatureInterpretation
- Signature Interpretation Analysis supports all analyses performed using the Interpretation matrix: https://github.com/macintyrelab/SignatureInterpretation_analyses
- Cell line preparation details how the raw cell line data was processed to be used for the analyses covered in Figure 3: https://github.com/markowetzlab/cell-line-signatures
- Bin size estimation details how the bin size was calculated that has been used to downsample the WGS to shallow WGS data: https://

github.com/markowetzlab/swgs-binsize

For the majority of analyses the free statistical software R (v3.6.1) was used. The following packages were used:
broom.mixed 0.3.3
caret 6.0-83
cmprsk 2.2-10
coin 1.4-2
compositions 2.0-1
cowplot 1.0.0
data.table 1.12.8
DNAcopy 1.64.0
flexmix 2.3-15
ggExtra 0.9
ggplot2 3.1.1
ggrastr 0.2.3
ggrepel 0.8.2
ggthemes 4.2.0
ggstance 0.3.3
ggpubr 0.2.5
gridExtra 2.3
lemon 0.4.4
jtools 2.0.3
lsa 0.73.1
MASS 7.3_51.5
NMF 0.21.0
patchwork 1.0.0
PharmacoGx 1.17.1
Polychrome 1.3.1
RColorBrewer 1.1-2
reshape2 1.4.3
rstatix 0.7.0
scales 1.0.0
sfsmisc 1.1-7
sjPlot 2.8.9
stringr 1.4.0
survival 3.1-11
survminer 0.4.8
tidyverse 1.3.1
vcfR 1.12.0
viridisLite 0.3.0
YAPSA 1.12.0

SignatureAnalyzer written in Python (cloned on 15/03/19, link: https://github.com/broadinstitute/SignatureAnalyzer-GPU) was used for the Signature discovery.

Data analysis

Signature derivation
Copy number profiles and absolute copy numbers produced with Affymetric Power Tools (v2.11.2) and ASCAT (R package, v2.4)
Mixture modelling: Flexmix (R package, v2.3-15); Variational Bayes Gaussian Mixture Modeling in Python (coded according to Blei and Jordan, Bayesian Anal., 2006)
Non-negative matrix factorisation: SignatureAnalyzer (Python script; Tan and Fevotte, IEEE Trans. Pattern Anal. Mach. Intell., 2013; Kim et al. Nature Genetics, 2016; cloned from Github on 15/03/19)
Signature assignment: LCD function from YAPSA (R package available from Bioconductor, v1.12.0)
CNA estimation from signature activities: lm function from base R (v3.6.1)

Signature robustness
ASCAT (R package, v2.4), ASCAT.sc (R package, unpublished)
Monte-Carlo simulation: code available in Github repository
Gaussian mixture modeling: Mclust function from mclust (R package, v5.4.6)

Simulations of cancer genomes, SBS signatures, single-feature vs multi-feature approaches
NMF (R package, v0.21.0)
SigProfilerExtractor (Python, v1.1.4)

Association with biological covariates AND drug response and drug target analysis
Z-score transformation: scale function from base R (v3.6.1)
Association of signature activities with whole-genome duplication: prop.test (Proportion test) and t.test (Welsh's t-test) function from base R (v3.6.1)
Difference in distributions: t.test function from base R (v3.6.1)
Cosine correlations: cosine function from LSA (R package, v073.1)
Spearman and Kendall correlations: cor and cor.test functions from base R (v3.6.1)
Permutation test: code available in Github repository
Cosine permutation test: cosinePerm function from PharmacoGx (R package, v1.17.1)
Linear and robust linear regression: lm function from base R (v3.6.1), rlm from MASS (R package, v73._51.5)
Wald test for testing coefficients of robust linear regressions: f.robftest function from sfsmisc (R package, v.1.1-7)
Fisher's exact test for count data: fisher.test function from base R (v3.6.1)
Hierarchical clustering: hclust and dist functions from base R (v.3.6.1)

p-value correction: p.adjust function from base R (v3.6.1) correcting for false discovery rate (Benjamini and Hochberg, J R Stast. Soc., 1995)

Survival analysis
Univariate survival analysis (Kaplan-Meier estimates): survfit function from survival (R package, v.3.1-11)
Multivariate survival analysis (Cox proportional hazard models): coxph function from survival (R package, v.3.1-11)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in this study were obtained from publicly available sources and are described in detail in Supplementary Table 1, section "Data and Code" in the Supplementary Methods. Some raw data have restricted access (TCGA dbGaP Accession number: phs000178.v11.p8; ICGC EGA Accession number: EGAS00001001692). Granting access to these is beyond the control of the authors. Access can be obtained by applying to the relevant Data Access Committees (see below for details). The authors declare that all other data supporting the findings of this study, including the source data for all figures, are publicly available without restrictions and also available from Supplement and the Github repositories.

Restricted data

- TCGA SNP6 Affymetrix arrays and copy number profiles
Paper: Beroukhim et al. (Nature, 2010)
URL to data:
TCGA Research Network: https://www.cancer.gov/tcga
Raw data: https://portal.gdc.cancer.gov/legacy-archive/search/f
Copy number segments: https://portal.gdc.cancer.gov/repository
Data access:
Part controlled (CEL files, raw data): Study Accession number (dbGaP): phs000178.v11.p8
Part open (copy number segments, processed data): ASCAT Github repository: https://github.com/VanLoo-lab/ascat/tree/master/ReleasedData

- PCAWG Copy number calls and Structural variants
Paper: Gerstung et al., Nature, 2020; Dentro et al., Cell, 2021
URL to data:
US TCGA portion of PCAWG: https://portal.gdc.cancer.gov/repository
ICGC portion of PCAWG: https://dcc.icgc.org/pcawg
Data access:
Controlled (raw data):
Study Accession numbers:
For US TCGA portion of the PCAWG data: Phs000178.v11.p8 (dbGaP) -> https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8#restricted-access
For the ICGC portion of the PCAWG data: EGAS00001001692 (EGA) with an application to the ICGC-DACO -> https://daco.icgc.org/

Databases
- The COSMIC database can be accessed with a free account without an application: https://cancer.sanger.ac.uk/cosmic

- The Reactome database is publicly accessible without an account: https://reactome.org/

- A snapshot of the Mitelman database can be accessed by the publicly available paper from Weaver and Cleveland: https://www.sciencedirect.com/science/article/pii/S0955067406001554?via%3Dihub A file summarising the relevant table 1 in the paper is found in our Github repository: https://github.com/markowetzlab/Drews2022_CIN_Compendium/blob/main/Section%201%20-%204%20Signature%20Discovery/input/external/WeaverCleveland2006_Table1.txt

Unrestricted public data

TCGA
- Affymetrix SNP6.0 arrays: Beroukhim et al. (Nature, 2010)
- Gene expression: COSMIC database (Tate et al., NAR, 2019)
- Somatic exon SNVs in TCGA: Ellrott et al. (Cell Systems, 2018)
- Clinical data: TCGA Research Network (https://www.cancer.gov/tcga)
- Germline and somatic BRCA1/2 status for OV and BRCA patients: Wang et al. (Nature Genetics, 2017); Maxwell et al. (Nature Communication, 2017)
- Promoter hypermethylation of RAD51C and BRCA1 in TCGA-OV samples: TCGA Research Network (Nature, 2011); Sample IDs obtained from corresponding author.
- Telomere length, TelomereSignatureScore, Genomic status and expression of TERT/TERC, TERRA expression, Alternative Telomere Lengthening (ATRX/DAXX): Barthel et al. (Nature Genetics, 2017)
- Driver genes and annotations: Bailey et al. (Cell, 2018)
- Extrachromosomal DNA amplicons: Kim et al. (Nature Genetics, 2020)
- Tandem duplication identification and classification: Menghi et al., (Cancer Cell, 2018)
- HRD score, telomeric imbalance score, large-scale state transitions score, loss of heterozygosity score: Knijnenburg et al. (Cell Reports, 2018), Marquard et al. (Biomarker Research, 2015)

- Cell cycle score (CCS): Lundberg et al. (Oncogene, 2020)

PCAWG
- Clinical data: ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (Nature, 2020) hosted on synapse.org (ID: syn10389158)
- SV breakpoints and copy number calls: Gerstung et al. (Nature, 2020); Dentro et al. (Cell, 2021)
- Structural variants and microhomologies: Li et al. (Nature, 2020)
- Chromothripsis data: ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (Nature, 2020)
- Kataegis data: ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (Nature, 2020)
- HRDetect predictions: Degasperi et al. (Nature Cancer, 2020)
- Exposures to SBS, DBS and ID signatures: COSMIC database (Tate et al., NAR, 2019)
- Exposures to Rearrangement signatures: Nik-Zainal et al. (Nature, 2016)

ICGC
- Breast cancer cohort plus clinical data: Nik-Zainal et al. (Nature, 2016)
- HRDetect predictions: Davies et al. (Nature Medicine, 2017)

Cell lines
- DepMap CRISPR screen: Meyers et al. (Nature Genetics, 2017)
- Broad institute cancer cell line encyclopedia (CCLE): Ghandi et al. (Nature, 2019)
- PRISM drug repurposing screen: Corsello et al. (bioRxiv, 2019)
- DEMETER2 RNAi screen: McFarland et al. (Nature Communications, 2018)

SNP loci: 31 downloaded from GDC (https://api.gdc.cancer.gov/data/7f0d3ab9-8bef-4e3b-928a-6090caae885b)

Pathway data and annotations: Reactome (Fabregat, PLoS Comp Bio, 2018; Jassal, NAR, 2020)

Ovarian copy number signatures: Macintyre et al. (Nature Genetics, 2018)

Public but unpublished data

The identifiers of the TCGA samples with RAD51C and BRCA1 promoter hypermethylation and TCGA-wide unweighted chromosome arm, whole-chromosome and focal CNA scores were published but not part of their respective papers or supplement (TCGA Research Network, Nature, 2011; Davoli et al., Science, 2017). We obtained the data from the corresponding authors and added them to the Supplementary Tables and the Github repositories.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The largest publicly accessible cancer cohort, the TCGA, was chosen as the main study cohort. From the initial 12,240 Affymetrix SNP6.0 arrays from 11,034 patients, of which 7,880 arrays for 7,880 patients (names "samples" from here on) fit the study design and passed computational quality control.

All other data sets, including the PCAWG (1898 patients) and the ICGC Breast cancer cohort (560 patients), were included to validate the findings on the TCGA cohort.

Data exclusions

We also provide this information in the Methods section of the manuscript.

The initial set of arrays included 11,034 patients with 12,240 samples. First, we excluded noisy samples by filtering for the following three criteria:
- Samples without an ASCAT solution (n = 99).
- Samples with ASCAT failures (n = 1,112). A purity equal to 100% generally indicates that ASCAT has identified a copy number profile from a normal tissue and therefore only samples with purity < 100% were considered.
- Samples with purity < 40% (n = 2,199). Samples with a low purity were removed to avoid low quality absolute copy number estimates.
For the resulting 526 multisample patients with 1,255 samples, we identified a representative sample based on the following criteria:
- sample passed quality control;
- sample is from the primary tumour;
- had blood as normal control;
- if there were still multiple options at that point, one was chosen at random.

Finally, the following three criteria were used for the exclusion of low quality arrays (n = 189):
- Tumour or normal MAPD < 0.75: The median of absolute pairwise differences between adjacent probes (MAPD) is a per-microarray estimate of the variance, similar to the standard deviation or the interquartile range36. Filtering for the MAPD removed noisy samples.
- Tumour and normal MAPD < 0.4 and fraction homozygous segments < 0.1: Occasionally tumour samples are matched with the wrong

germline sample. This results in a high fraction of homozygous segments. The above thresholds removed these samples.
- Differences between logR and BAF segments < 250: We identified and removed 67 samples from 17 cancer types with extreme noise in logR values, which manifested as a large-scale wave pattern which could not be captured by the median of absolute pairwise differences between adjacent probes (MAPD). The corresponding B-allele frequencies (BAF) values did not show this wave pattern and generally appeared consistent. These samples were identified by having more than 250 segments difference between their logR and BAF segmentations.

All filters combined resulted in 7,880 high-quality copy number profiles across 33 cancer types (Extended Data Fig. 1a).

Afterwards, we developed a method to identify patients with detectable chromosomal instability which excluded another 1,545 samples, resulting in the final study cohort of 6,335 samples.

Replication

The signatures were produced from 6,335 samples from 33 cancer types with detectable chromosomal instability. To our knowledge there is no other pan-cancer SNP6.0 cohort available with a similar sample composition.

We validated our approach by simulating cancer genomes shaped by five well-studied mutational processes (Supplementary Figs. 3-6, Methods). In a second simulation study we derived signature-specific activity thresholds, tested the stability of signature definitions, and the stability of signature activities (Methods, Extended Data Fig. 5, Supplementary Fig. 7). We then tested the robustness of our approach across different high-throughput technologies using 478 samples profiled by both TCGA (SNP 6.0 and whole exome sequencing, WES) and PCAWG (whole genome sequencing, WGS). We compared signature definitions and activities across five platforms: SNP 6.0 without matched normal, WGS downsampled to SNP 6.0 positions, WGS downsampled to shallow WGS, on-target WES and off-target WES. Quantifying the activity of our signatures was robust across all platforms. Re-deriving the signatures was possible across WGS platforms but performance deteriorated for WES (Extended Data Fig. 6).

Randomization

No randomisation was performed - this was a descriptive study, not an experimental study.

Blinding

No blinding was undertaken - this was a descriptive study, not an experimental study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study

☒ □ Antibodies
☒ □ Eukaryotic cell lines
☒ □ Palaeontology and archaeology
☒ □ Animals and other organisms
☒ □ Human research participants
☒ □ Clinical data
☒ □ Dual use research of concern

Methods

n/a | Involved in the study

☒ □ ChIP-seq
☒ □ Flow cytometry
☒ □ MRI-based neuroimaging