
Supplementary information

A pan-cancer compendium of chromosomal instability

In the format provided by the
authors and unedited

Supplementary information

A pan-cancer compendium of chromosomal instability

In the format provided by the
authors and unedited

Supplementary Methods

Data and code	4
Code availability	4
Data sources and availability	4
Databases	5
Generating copy number profiles	10
Inferring absolute copy number from SNP 6.0 arrays	10
Identifying high-quality copy number profiles	10
Computing detectable chromosomal instability (dCIN)	11
Copy number feature encoding	13
Feature encoding	14
Segment size	14
Changepoint	14
Lengths of chains of oscillating copy numbers	14
Breakpoints per 10MB and chromosome arm	14
Additional details	15
Exclusion of absolute copy number as a feature	15
Exclusion of loss of heterozygosity (LOH) as a feature	16
Feature coverage of known CIN patterns	16
Feature distributions and mixture modelling	17
Deriving signatures	19
Choosing a matrix-decomposition method	19
Deriving pan-cancer signatures	20
Deriving cancer-type enriched signatures	20
Merging to a signature compendium	20
Estimating the number of CNAs per signature	21
Assessing feature encoding performance	22
Simulating CIN	22
Creating gold-standard signatures	25
Identifying simulated signatures	25
Single-count versus multi-count feature encoding	26
Robustness analysis	28
Signature stability assessment via simulation	28
Signature activity stability	28
Signature definition stability	28
Signature-specific thresholding	29
Signature stability across genomic technologies	29
Step 1: Absolute copy number fitting	29
Approach 1: SNP 6.0 without matched normals	29
Approach 2: WGS downsampled to SNP 6.0 positions	30

Approach 3/5: Bin size estimation allowing for downsampling to shallow WGS	30
Approach 3: Shallow WGS: Absolute copy number fitting	31
Approach 4: WES: Absolute copy number fitting of on-target reads	32
Approach 5: WES: Absolute copy number fitting of off-target reads	32
Step 2: Extract signature activities and definitions	33
Step 3: Comparison to gold standard (SNP6 with matched normal)	33
Recommendations for WGS and WES data	33
Interpreting signature definitions	35
Signatures versus mutational processes	35
Feature shrinkage is a general property of NMF	35
The signature interpretation matrix	36
Validation of the signature interpretation matrix	37
Identification of putative signature aetiologies	38
Linking signature activities to mutated genes	38
Selection of 1,146 genes of interest	38
BRCA1, BRCA2 and RAD51C mutation or hypermethylation	39
Testing for differences in signature activity	40
Selecting high-confidence gene associations	41
Correcting for extrachromosomal DNA amplicons	41
Correcting for cancer type	42
Proposing an aetiology and assigning a star rating	42
Focussed analysis of CCNE1 and CDK12	43
Supporting putative aetiologies with orthogonal data	44
Chromothripsis	45
Kataegis	46
Loss of heterozygosity (LOH)	46
SBS, DBS and ID signatures	46
Rearrangement (SV) signatures	46
Patient age	46
Whole-chromosome CNAs	47
Whole-genome duplication status and genome instability score	47
Identification of signatures related to ploidy changes	48
Telomere length and elongation	48
Tandem duplications	49
Cell cycle score	49
Centrosome amplification	49
Seven HRD metrics	50
Microhomologies of PCAWG structural variants	50
Ovarian copy number signatures	51
Gene expression of DNA repair genes	51
Ethnicity	52
Mitotic catastrophe	52
Impaired HR signatures on breast and ovarian cancer samples	53

ecDNA	53
Testing for independence between CX2, CX3 and CX5	54
Correlating signature activity with drug response	55
Calculating signature activities for cell lines	55
CRISPR, RNAi and drug response data integration	55
Correlation with target gene KO, KD and drug inhibition	56
Building a clinical classifier based on signature activities	57
Preparation of survival data	57
Survival analysis	57
Clinical classifier based on CX2/CX3 signature activity	57
Comparison with HRDetect and Myriad myChoice classifier	58
Clinical classifier based on three impaired HR signatures	58
Logistic regression on survival outcome	58
Support vector machine (svm) on survival outcome	58
Support vector machine (svm) on hierarchical clustering	58
Reference genome	59
Software versions	59
References	60

Data and code

Code availability

The code is publicly accessible via the hub repository <https://github.com/markowetzlab/Dreus2022>, 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 specialised tasks;

Code repositories

- [Dreus2022 CIN Compendium](#) is the main github repository containing code to reproduce the figures, most results and most supplementary tables
- [CINSignatureQuantification](#) covers the R package needed to quantify signature activities from copy number profiles. Also includes clinical classifier
- [CINSignatureDiscovery](#) covers 1) how the signature compendium was obtained and 2) details how to identify de-novo signatures from copy number profiles
- [CINSignatureGenomeSimulation](#) covers the simulation we used to validate the multi-feature approach
- [CINSignatureBiomarkerAnalysis](#) supports the results shown in Figure 3
- [CINSignatureInterpretationMatrix](#) details how the Interpretation matrix was obtained
- [CINSignatureInterpretationAnalysis](#) supports all analyses performed using the Interpretation matrix
- [CINSignatureCellLines](#) details how the raw cell line data was processed to be used for the analyses covered in Figure 3
- [CINSignatureSWGSBinSize \(Shallow WGS bin size estimation\)](#) details how the bin size was calculated that has been used to downsample the WGS to shallow WGS data
- [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

We have also made our results available as an online resource at <https://markowetz.cruk.cam.ac.uk/cincompendium/> for users to interact with and explore the data.

This level of data and code sharing is Bronze standard of reproducibility⁵¹.

Data sources and availability

Supplementary Table 1 lists the data used in this study and their sources. All data are publically accessible, with restrictions for the TCGA, PCAWG and ICGC raw data (e.g. SNP 6.0 arrays, WGS, WES). Access to the TCGA data (SNP 6.0 arrays, WES) and the TCGA-part of the PCAWG data (WGS) can be obtained by a Data Access Request through the Database of Genotypes and Phenotypes (dbGaP, [Link](#)). For the ICGC data (SNP 6.0)

and non-TCGA PCAWG data (WGS), an application to the ICGC DACO ([Link](#)) has to be made. We are not responsible for maintaining or controlling access to the data.

Databases

The COSMIC database can be accessed with a free account without an application ([Link](#)). The Reactome database is publicly accessible without an account ([Link](#)). A snapshot of the Mitelman database can be accessed by the publicly available paper from Weaver and Cleveland ([Link](#)). A file summarising the relevant table 1 in the paper is found in our github repository ([Link](#)).

Supplementary Table 1: Sources of publicly available data

Data	Source
TCGA SNP6 Affymetrix arrays and copy number profiles	Paper: Beroukhim et al.⁵² URL to data: • TCGA Research Network • Raw data • Copy number segments Data access: • Part controlled (CEL files, raw data) • Part open (copy number segments, processed data) Alternative access: ASCAT Github repository Study Accession number (dbGaP): phs000178.v11.p8
TCGA Gene expression	Paper: Beroukhim et al.⁵² URL to data: COSMIC⁵³ Data access: open
PCAWG Copy number calls and Structural variants	Paper: Gerstung et al. and Dentre et al.^{54,55} URL to data: • US TCGA portion of PCAWG • ICGC portion of PCAWG Data access: • Part controlled (raw data) • Part open (copy number segments, processed data) Study Accession numbers: For US TCGA portion of the PCAWG data: • Phs000178.v11.p8 (dbGaP) For the ICGC portion of the PCAWG data: • EGAS00001001692 (EGA) with an application to the ICGC-DACO

DepMap CRISPR screen	Paper: Meyers et al.⁵⁶ URL to data: https://depmap.org/portal/ Data access: open Name of the file: CRISPR_gene_effect.csv
Cell lines from the CCLE project	Paper: Gandhi et al. 2017⁵⁷ URL to data: https://depmap.org/portal/ Data access: open
PRISM drug repurposing screen	Paper: Corsello et al.⁵⁸ URL to data: https://depmap.org/portal/ Data access: open Name of the file: • secondary-screen-dose-response-curve-parameters.csv • secondary-screen-cell-line-info.csv
DEMETER2 RNAi screen	Paper: McFarland et al.⁵⁹ URL to data: https://depmap.org/portal/ Data access: open Name of the file: D2_combined_gene_dep_scores.csv DEMETER_sample_info.csv
Germline and somatic BRCA1/2 status for TCGA-OV and TCGA-BRCA patients	Paper: Wang et al. and Maxwell et al.^{60,61} URL to data and name of the file: Supplementary Table 13 (Wang et al.) Supplementary Data 1 (Maxwell et al.) Supplementary Data 2 (Maxwell et al.)
Promoter hypermethylation of RAD51C and BRCA1 in TCGA-OV samples	Paper: TCGA²⁰ URL to data: TCGA2011 BRCA1 RAD51C Calls 20100718.txt Note: Sample IDs not published in original paper but have been obtained from Paul Spellman, corresponding author. Data access: open
TCGA clinical data (incl. survival data)	Paper: TCGA⁶² URL to data and name of the file: clinical.cases_selection.<date>.tar.gz Data access: open
PCAWG clinical data (incl. survival data)	Paper: PCAWG⁶³ URL to data: Hosted on synapse.org

	Accession number: syn10389158 Name of the file: pcawg_donor_clinical_August2016_v9.xlsx Data access: open
ICGC 560 breast cancer samples plus clinical data	Paper: Nik-Zainal et al.²¹ URL to data: https://www.nature.com/articles/nature17676#Sec22 Data access: open
Triple-negative status for TCGA breast cancer samples	Paper: Lehmann et al.⁶⁴ URL to data and name of the file: pone.0157368.s008.xlsx Data access: open
extrachromosomal DNA for TCGA	Paper: Kim et al.⁶⁵ URL to data and name of the file: Supplementary Tables Data access: open
Tandem duplication classification for TCGA	Paper: Menghi et al.¹⁷ URL to data and name of the file: Table S3 Data access: open
HRDetect on 560 breast cancers	Paper: Davies et al.⁶ URL to data and name of the file: Supplementary Table 3 Data access: open
HRDetect on PCAWG	Paper: Degasperi et al.⁶⁶ URL to data and name of the file: Supplementary Tables Data access: open
HRD score, telomeric imbalance score, large-scale state transitions score, loss of heterozygosity score	Paper: Knijnenburg et al. and Marquard et al.^{25,67} URL to data and name of the file: Table S1 Download from Gerke lab github Data access: open
Cell cycle score (CCS)	Paper: Lundberg et al.⁶⁸ URL to data: https://github.com/arianlundberg/PANCAN.analysis Note: Please follow the script “Figure1.R” to extract the TCGA-specific cell cycle scores.

	Data access: open
Somatic exon SNVs in TCGA	Paper: Ellrott et al. ⁶⁹ URL to data and name of the file: mc3.v0.2.8.PUBLIC.maf.gz Data access: open
Exposures to SBS, DBS and ID signatures	Paper: Alexandrov et al. ³⁸ URL to data: Explained on COSMIC website ⁵³ and hosted on synapse.org Accession number: syn12009743
Exposures to rearrangement (SV) signatures	Paper: Nik-Zainal et al. ²¹ URL to data: https://www.nature.com/articles/nature17676#Sec22 Data access: open
Chromothripsis and Kataegis	Paper: PCAWG et al. ⁶³ URL to data: https://dcc.icgc.org/releases/PCAWG/evolution_and_heterogeneity/clustered_mut_processes Data access: closed, Please apply to the ICGC DACO to gain access (https://daco.icgc.org/).
Telomere length, TSS, TERT/TERC, TERRA, Alternative Telomere Lengthening (ATRX/DAXX)	Paper: Barthel et al. ⁷⁰ URL to data and name of the file: Supplementary Table 1 Data access: open
Driver gene classification	Paper: Bailey et al. and Jassal et al. ^{71,72} URL to data and name of the file: Document S1. Tables S1–S7 Note: Other driver gene classifications were downloaded from the Reactome database. Please see the section “Linking signature activities to mutated genes” for more details. Data access: open
SignatureAnalyzer aka ARD-NMF (used for signature derivation)	Paper: Tan and Fevotte, Kim et al. and Alexandrov et al. ^{38,73,74} URL to github repo: https://github.com/broadinstitute/SignatureAnalyzer-GPU Data access: open
Other detectable CIN measures	Paper: Weaver and Cleveland, Davoli et al. and Taylor et al. ^{75–77}

	Note on Weaver and Cleveland: This paper was used to get a snapshot of the Mitelman database. It can be justified as this paper is still widely cited. Please see the section “Computing detectable chromosomal instability (dCIN)” for more information. Note on Davoli et al: The relevant unweighted arm, chromosome and focal scores were obtained by directly contacting the corresponding authors. The file can be downloaded from our github repository Davoli2017_Supplement_from_author.txt URL to data and name of the file (Taylor et al): Table S2. Sample Chromosome Arm Calls and Aneuploidy Scores, Related to Figure 1 Data access: open
SNP loci from the 1,000 Genome Project (phase 3) falling into exonic regions (GENCODE v19 exome)	Paper: Frankish et al.⁷⁸ URL to data: GDC Data access: open

Generating copy number profiles

Inferring absolute copy number from SNP 6.0 arrays

We processed sample-specific CEL files from TCGA with Affymetrix Power Tools (APT; v2.11.2; standard options) and ASCAT⁷⁹ (R package, v2.4). By default ASCAT produces an absolute, genome-wide copy number segmentation where segments are rounded to the nearest integer copy number state and merged. However, our copy number signature framework required unrounded copy number segments in order to avoid artificially removing signal (e.g. subclonal copy number changes). We determined an optimal segmentation penalty to avoid over- or under-segmentation by comparing ASCAT generated breakpoints with multiple segmentation penalties (25, 35, 50, 70, 100, 140, 200) to the PCAWG consensus copy number calls for 714 samples from 22 cancer types which were generated using whole-genome sequencing. We found that a penalty of 70 offered the highest degree of overlap (**Supplementary Fig. 39**). To remove any noise around diploid copy states, we collapsed all segments around 2 ($1.9 < x < 2.1$) to exactly 2 and merged adjacent segments. This merging was also performed for deleted segments (copy number equals 0).

Identifying high-quality copy number profiles

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:

- 1) Samples without an ASCAT solution ($n = 99$).
- 2) 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.
- 3) 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:

- 1) sample passed quality control;
- 2) sample is from the primary tumour;
- 3) had blood as normal control;
- 4) 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 ($n = 189$):

- 1) 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 range⁸⁰. Filtering for the MAPD removed noisy samples.
- 2) 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.

- 3) 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**).

Computing detectable chromosomal instability (dCIN)

For signature identification, we aimed to use only samples with sufficient evidence of chromosomal instability (CIN) and established a pragmatic threshold for the detectability of CIN (dCIN) from copy number profiles. Our approach uses our previous ovarian cancer signature method⁸ to estimate which CNAs in a sample can be robustly assigned as coming from a CIN signature and which CNAs cannot. The number of residual, unassigned CNAs can be used as an indicator of the minimum number of CNAs required to detect a signature of CIN, and thus establish an empirical threshold for detectable CIN.

We started by calling ovarian copy number signatures on 539 high-grade serous ovarian cancer samples⁸ using the method described in Macintyre et al.⁸ to generate the input matrix, and YAPSA⁸¹ (R package, v1.12.0) and the signature definition matrix from Macintyre et al.⁸ to call signatures. We considered signatures with activity above 5% in these samples to be robustly detected signatures of CIN and considered signatures below 5% as potentially representing residual CNAs. To convert these <5% signature activities into an estimate of the number of assigned CNAs, we set signature activities above 5% in the activity matrix to 0 (leaving only the signal from the signatures with less than 5% activity), and multiplied this matrix by the signature definition matrix to derive an input matrix containing signal which representing residual CNAs which could not be robustly assigned to any signature. For each sample and feature, we summed the remaining probabilities to estimate the number of residual CNAs. For each feature, we defined the quantile 0.95 of the distribution of residual CNAs. From this distribution of quantiles 0.95, we took the 3rd quartile (which is less volatile and more robust than the maximum) as the empirical threshold of the detectability of CIN which was 20 CNAs (**Extended Data Fig. 1b**).

Previous studies to quantify the proportion of cancer samples with CIN used aneuploidy as a surrogate measure with pan-cancer estimates ranging between 60% and 95% depending on the method (karyograms⁸²; genomics case studies^{83,84}; computational methods⁸⁵; FISH⁸⁶; reviewed¹⁰). Multiple studies found that around 90% of solid tumours are aneuploid^{5,10,75,76,82,83,87}. This number goes back to a snapshot of the Mitelman database of chromosome aberrations and gene fusions⁷⁵, which analysed 17 solid and 9 hematopoietic cancers of which 13 could be matched to organs included in the TCGA project (**Extended Data Fig. 1c**). Apart from LAML (acute myeloid leukemia) and UVM (uveal melanoma), all cancer types agreed with minor discrepancies of up to 20%. The larger discrepancies for UVM can be explained as it was mapped onto retinoblastoma, two fairly different diseases⁸⁸. For LAML, the majority of samples had a near diploid number of chromosomes (<68)⁷⁵.

Given that LAML samples have quiet copy number landscapes⁴⁹, it is most likely the case that the dCIN threshold with 20 CNAs lies just above the average changes in LAML.

We compared our dCIN measure to two other measures previously published in the literature: the aneuploidy score from Taylor et al.⁷⁶ and SCNA levels from Davoli et al.⁷⁷. Taylor and colleagues analysed 10,522 SNP arrays from the TCGA and found 88% aneuploid samples. Davoli et al. did not summarise a single percentage of how many cancers have aneuploidy, instead they supplied a single somatic CNA (SCNA) score plus three more specific scores for whole chromosomes, arms and focal changes. We divided the samples into samples with detectable CIN and undetectable CIN and compared the measures from Davoli and Taylor. All comparisons were highly significant (**Extended Data Fig. 1**). As before, the dCIN threshold missed some samples with mostly quiet profiles largely containing whole chromosome CNAs and arm level changes. It is unclear whether such quiet genomes were produced by a mutational process or simply originated from random mitotic errors and unrepaired double-strand breaks.

Copy number feature encoding

Modelling the effects of different types of chromosomal instability presents a unique challenge compared to signatures of single-nucleotide variants (SNV) and structural variants (SV). This challenge stems from the fact that for CNA analyses we need to distinguish between a genome-changing event and the number of CNAs it produces. We consider an event to be the result of a *single* error in a regular biological process which leads to chromosomal instability (e.g an error during mitosis). The resulting event may involve one or more copy number changes depending on the type of error. For instance, a single whole chromosome missegregation event will result in a single copy number change. Whereas, a single chromothripsis event will result in many CNAs. Therefore, to adequately model the activity of different types of CIN, we need to model activity in terms of numbers of events, not numbers of CNAs, and the number of CNAs needs to be decoupled from the number of events. This is different from SNV signatures, where the number of SNVs is assumed to be a good measure of the activity of a mutational process.

Developers of SV signatures faced a similar problem and Li et al.⁸⁹ addressed it by performing extensive clustering of SVs into mutually exclusive “event” types, ranging from events that encompassed a single deletion of a certain size, to events that involved large numbers of SVs as part of chromoplexy chains. This facilitated the counting of these events in a tumour and the application of NMF, similar to what had been previously done for SNVs. The approach of Li et al thus allowed SV signatures to be interpreted in a similar manner to SNV signatures.

Copy number data, however, presents unique challenges, precluding the use of a similar approach. CNAs cannot be easily clustered into single “events” as is the case for SVs, as they lack the information required to relate one CNA to another. Central to this problem is the fact that copy number changes are quantified with respect to reference genome coordinates. As a result, for example, the loss of a “chimeric” chromosome (an already rearranged chromosome containing parts of many different chromosomes), is a single “event” in the tumour cell; but, mapped to a reference genome, this event can manifest as many copy number changes across multiple chromosomes which are not linked to each other. Therefore, it is not possible to generate a vector of “event” counts and thus approaches similar to those used in the SNV and SV setting cannot be applied.

Modelling copy number signatures demands a completely new and different approach. As originally published in Macintyre et al.⁸ and substantially enhanced here, we used the published literature to define a set of copy number related features representing known outcomes of different types of CIN including chromothripsis, tandem duplication, whole-genome duplication and focal amplification (**Supplementary Table 2**). This formed the basis of our set of 5 features.

Embedding CNAs observed in a tumour into this feature space, means single CNAs can add weight to multiple features. This has the significant advantage of allowing “events” to be encoded that correspond to low or high numbers of CNAs, and thus overcomes the significant challenges outlined above of working with copy number data (see section **Single-**

versus multi-feature encoding). However, this approach does have some potential caveats when interpreting the signature definitions (see section **Signature interpretation matrix**).

Feature encoding

Segment size

The segment size feature refers to the length in base pairs of copy number aberrations, which are segments defined by a copy number unequal to 2 ($x < 1.9$ or $x > 2.1$). Diploid segments were not considered.

Changepoint

The changepoint feature captures the difference in absolute copy number compared to the left neighbouring segment. For diploid segments no changepoint value is extracted. For non-diploid segments, the changepoint value is calculated with respect to the segment to the left (even if the neighbouring segment is diploid). If the first segment of a chromosome is a CNA, then an artificial diploid segment is used to calculate a changepoint value for this CNA, i.e. two is subtracted from the absolute copy number value. Capturing copy number this way avoids a conflation by the background ploidy status (for more details see section “Exclusion of absolute copy number as a feature”, **Methods page 15**).

Lengths of chains of oscillating copy numbers

Certain mutational events can lead to intermittent losses (or gains) of segments, leading to an oscillating change in copy number states. One example is chromothripsis, where a chromosome breaks up in multiple pieces and not all pieces are stitched together by DNA repair pathways, which can then ultimately be lost. The resulting pattern is an oscillating change between two copy number states. This feature captures the length of such chains by rounding the copy number states to the next integer state and then counting how many times a chain is changing between the state of the first segment and another copy number state.

Breakpoints per 10MB and chromosome arm

We included two features capturing potential clusters of CNAs: breakpoints per 10 Mb and breakpoints per chromosome arm. These two features are needed to accurately model events of different sizes: one for small and medium sized clusters with the largest possible window size and one feature for larger clusters. In addition, the breakpoints per chromosome arm feature allows chromosome-specific biology to be taken into account.

Clusters of CNAs come in various sizes, e.g. short tandem duplications smaller than 10 kb length, long tandem duplications with lengths of over 100 kb each, and large-scale transitions and chromothriptic events which can span multiple dozens of mega bases^{6,17,19,90}. Their respective cluster sizes can be in the tens of kb up to whole chromosome arms. However, the smallest chromosome arm on a SNP 6.0 array is 12.8 Mb in length. Having larger window sizes than 10 Mb would therefore lead to a skewed breakpoint density on smaller chromosome arms. A smaller window size might be unfeasible to capture medium sized clusters, which would require introducing a third feature capturing clusters of CNAs.

Additional details

We collapsed and merged near diploid segments to a diploid state ($1.9 < x < 2.1$). During feature extraction, we then ignored these diploid segments for the segment size and the changepoint feature to avoid inflating these distributions in genomes with moderate CIN.

These changes had overall little impact on the ovarian signature activities from the TCGA ovarian cancer samples (**Supplementary Fig. 40**). The major difference was a shift between signatures 3 and 7, explained by the fact that signature 3 is acting on a diploid genome and 7 on a tetraploid genome, suggesting that both signatures represented the same mutational process artificially split up by the copy number feature.

Furthermore, it is of interest to know how the changes to the handling of diploid segments affected the relationship between samples which had highly similar signature activities. First, we studied sample TCGA-09-1673 which had the most CNAs (700) and segments (720) in this cohort (**Supplementary Fig. 41a**). The third most similar sample according to the cosine similarity between ovarian signature exposures (cosine similarity 0.936) was sample TCGA-24-2026 which had a more moderate CIN landscape with 354 segments and 353 CNAs. After removing diploid segments, the cosine similarity improved slightly to 0.941. Despite their different CIN landscape, both samples remained similar after removing diploid samples. Looking at the overall change in cosine similarities between signature activities before and after removing diploid segments shows that overall similarities increased slightly (**Supplementary Fig. 41b**). 95% of cosine similarities changed between 0.27 and -0.06 with a mean of 0.05 and a median of 0.029.

Diploid segments can either be normal segments or the result of copy-neutral LOH. Under the assumption that most diploid segments were normal segments, this improvement in cosine similarities is explained by removing the artificial inflation of segments, ultimately influencing multiple copy number features and thus signature activities. The increase in cosine similarity can therefore be seen as a refinement in signature activities. These findings also indicate that normal, diploid segments influenced signature activities even on strongly rearranged genomes and therefore should be removed from a pan-cancer setting where the majority of samples had quieter, less rearranged genomes.

Exclusion of absolute copy number as a feature

In a major deviation from our previous work⁸, we did not use the absolute copy number of a segment as a feature, because we found that it can artificially split a signature by ploidy status despite representing the same mutational process. As multiple copy number changes can happen at the same genomic location, the resulting absolute copy number of that segment can vary wildly depending on the order of events. For example, a single copy loss on a whole-genome duplicated background results in a copy number state of 3 but a deletion in a non-WGD background would result in a copy number state of 1 and therefore be encoded by two different signatures, even though they might be caused by the same mutational process that results in deletion patterns. This effect is present in the original ovarian copy number signatures⁸, where signature 3 (BRCA1/2-related homologous recombination deficiency) acts on a diploid genome and signature 7 (non-BRCA1/2-related homologous recombination deficiency) on a tetraploid genome. The rest of the signature

definitions are very similar. When we rederived the ovarian signatures without the copy number feature we found six signatures (rather than the published seven), with signatures 3 and 7 collapsed to one.

A few studies looked at BRCA1/2-related homologous recombination deficient cancer samples and tetraploid genome status^{19,91–93}. Jonsdottir et al. found that BRCA2-linked homologous recombination deficient tumours were three times more likely to have a tetraploid genome⁹¹. Two studies found no significant difference between HRD in tetraploid and diploid genomes⁹². However, a study on a small cohort of basal-like breast cancers found that BRCA1 germline mutations happen more often in diploid (12/24 samples, 50%) than in tetraploid cancers (11/41 samples, 27%)¹⁹. The more complex role of WGD arising from these publications supported our notion to exclude copy number as a feature and thereby correcting for background ploidy.

Exclusion of loss of heterozygosity (LOH) as a feature

We are not using a feature to describe LOH, because cheap genomic technologies like shallow WGS, which might be very useful in the clinic, do not have allele-specific information. Loss of heterozygosity (LOH) is the loss of one allele and is a common genetic event during cancer evolution and might pose a novel class of cancer vulnerabilities by removing genetic redundancy in cancer cells^{94,95}. Most LOH events are so-called copy-loss events and can be detected through a change in copy number. In our data set, the median length of LOH events across the TCGA was 3.6 million bp, which suggests that the majority of events can be detected by the SNP6 technology. Implementing a specific LOH feature, i.e. the proportion of the major allele over both alleles, would add the ability to detect copy-neutral LOH and probably add resolution to further differentiate on copy-loss LOH events. But this would require the use of a high-throughput method allowing for the differentiation of the signal between alleles, such as WGS or SNP 6.0 arrays. However, with the advent of shallow WGS and single-cell sequencing in the scientific mainstream which do not yet allow allele-specific resolution^{8,96–98}, a dedicated LOH feature might lead to constraints in the use of the future signatures.

Feature coverage of known CIN patterns

Supplementary Table 2 lists well-studied copy number patterns and describes which specific copy number feature is able to capture their unique characteristics. These observations motivate our selection of five fundamental features to model known patterns and potentially distinguish their mutational processes.

Supplementary Table 2: Copy number features capture known copy number patterns

Seg .size = Segment size, *CNC* = Copy number changepoint, *BP10* = Breakpoints per 10 Mb, *BPARM* = Breakpoints per chromosome arm, *OSC* = Length of oscillating copy number states

Pattern	Citation	Seg. size	CNC	BP10	BPARM	OSC
Amplification	99	x	x	x		
Aneuploidy	76	x	x			
Breakage-fusion-bridge	100,101		x		x	
Chromothripsis	90		x	x	x	x
Complex genomic rearrangement	102	x	x	x	x	x
Deletions	49		x			
Extrachromosomal DNA (ecDNA)	65,103,104	x	x			
Homologous recombination deficiency (HRD)	6,18,19,105	x	x	x	x	x
Loss of heterozygosity (LOH)	94	x	x		x	
Micronuclei	106	x	x			
Tandem duplications	17,107	x	x		x	x
Whole-genome duplication (WGD)	49,55		x			

Feature distributions and mixture modelling

After obtaining the distributions for the five features, we aimed to split them into categories (or components) that were supported by the data and included as few design choices as possible. Furthermore, we want to use the probability of each value to belong to each component in order to include uncertainty around the values. Applying mixture modeling to the feature distributions, we identified a core set of underlying copy number feature components.

For the segment size and changepoint feature distributions, being (quasi)-continuous data, we fitted Dirichlet-Process Gaussian mixture models using variational inference (code implementation followed Blei and Jordan¹⁰⁸, 128 initial components and we ran 10,000 iterations for segment size distribution, and 5,000 iterations for changepoint distribution), resulting in 30 and 22 mixture components, respectively, with weights of more than 1%. Mixture components were merged if a mean was within or near the first standard deviation of another mixture component. The mean of the newly merged component was calculated by

the weighted mean of the involved components. Standard deviation was estimated by sampling 100,000 points from the inferred distributions. This reduced the number of mixture components from 30 down to 22 for the segment size mixture model and from 22 down to 10 for the changepoint mixture model.

For the count distributed breakpoints per 10MB, breakpoints per chromosome arm and lengths of oscillating chains, we used Poisson mixture models from flexmix¹⁰⁹ (R package, v2.3-15) and used the Bayesian Information Criterion (BIC) to decide on the number of components. We derived 3 components for the breakpoints per 10Mb distribution, 5 for the breakpoints per chromosome arm, and 3 for the length of oscillating chains.

In total, we derived 43 mixture components. Once derived, we calculated the probability for each feature value to belong to each component. For a given sample, the probabilities of each mixture component were then summed together, resulting in a 1x43 dimensional vector of sum-of-posterior probabilities. Deriving the sum-of-posterior vector for all 6,335 samples resulted in a 6,335 by 43 sum-of-posterior matrix, referred to as input matrix from here on.

Deriving signatures

Choosing a matrix-decomposition method

Our feature encoding introduces substructure in the input matrix that is different to that of a typical SNV count matrix. This structure does not violate the only assumption of NMF that the matrix does not contain negative numbers. When faced with deciding on which approach to use for matrix decomposition, we identified two possibilities. 1) A bespoke approach, similar to the Hierarchical Dirichlet model with latent variables that was used in Li *et al.*⁸⁹, where we specifically take into account the substructure of our feature encoding. Or 2) the widely used NMF approach, which is agnostic to the substructure.

The advantage of the HDP-like approach is that we can impose prior knowledge that certain features should behave in certain, coordinated ways. However, as the feature encoding itself already strongly imposes prior knowledge, this additional constraint may suppress any ability to perform data driven discovery and could be overly biased by our prior knowledge.

The advantage of using NMF is that it is robust and widely used across many scientific domains, including face recognition, astronomical image analysis and text mining. It has well defined behaviour and provides a data driven approach to signature identification across our feature space.

In this work we aimed for a balance between using prior knowledge and data driven discovery: we encoded prior knowledge in the selection of copy number features (which were designed to capture known patterns of CIN) but used a widely used, well-tested, and robust “off the shelf” method for deconvolution.

There are multiple NMF algorithms successfully used in the biological literature^{74,110,111}. The use of the right NMF algorithm is subject to the shape of the data as well as more pragmatic considerations such as run time or memory requirements. Our input matrix consists of five different features. One approach would be to use a NMF that is able to work with multiple input matrices such as joint NMF¹¹². However, one of the constraints of joint NMF is that the number of possible identifiable signatures must be less than the minimum number of components observed for each feature (or sub matrix). This would mean, without substantially changing our features the maximum number of signatures that could be identified would be 2. Therefore, we used SignatureAnalyzer^{38,73,74} (Python 3.6.7, cloned on 15/03/19) NMF that works on one input matrix but was optimised for runtime by allowing for the computations to be outsourced to a GPU.

To compile a compendium of pan-cancer copy number signatures, we adapted the approach outlined in Macintyre et al.⁸ to identify signatures across all tumours, pan-cancer, and those which might be specific to the 33 different cancer types. We then developed a merge and prune procedure to combine these signatures into a robust and orthogonal compendium of pan-cancer copy number signatures. The following paragraphs explain these steps in detail.

Deriving pan-cancer signatures

We used non-negative matrix factorisation (NMF) on the input matrix. The 43 components contain copy number events generated by all types of CIN and NMF facilitates the identification of separate types of CIN.

We used SignatureAnalyzer with an L1 regularisation on both output matrices to perform non-negative matrix factorisation (NMF) on the input matrix. We ran the NMF 200 times with parameter K0 (maximum number of allowed signatures) set to default value, in our case 43. The NMF was initialised randomly to navigate the non convex optimisation function and compared the overall Kulback-Leibler (KL) divergence between the input matrix and the reconstructed input matrix from the output matrices. A lower KL divergence indicates a better solution. The number of signatures, K, was defined as the mode of the distribution of K's over the 200 runs, which was equal to 10 for all 6,335 samples pan-cancer.

Of the 200 runs, 95 solutions had a K of 10. To pick a representative solution, we first calculated the cosine similarity between the signatures in the 95 sets and visualised the results as a graph. The graph had 950 nodes, each one representing a signature, and edges connecting nodes if the cosine similarity between these two signatures was above 0.85, a threshold established in the literature³⁸. This graph showed 12 independent clusters with sporadic outliers attached to them, as well as a few single independent nodes (**Supplementary Fig. 42**). In a second step, we ordered the 95 solutions by their KL divergence and, starting with the highest-ranked solutions, evaluated their spread in the graph. We chose the 9th optimal solution as it had a signature in 10 of the 12 clusters. The previous 8 solutions all had at least two signatures in the same cluster, indicating almost identical signatures. In terms of the Kullback-Leibler divergence, the solution we picked is about 1% worse than the most optimal solution we derived.

Deriving cancer-type enriched signatures

Next, we aimed at deriving cancer-type enriched signatures to complement the pan-cancer set of 10 signatures. We hypothesised that there might be signatures present in a certain cancer type but that their signal may be drowned out during pan-cancer NMF. Therefore, for 20 of the 33 cancer types with over 100 samples with detectable CIN, we derived cancer-type enriched signatures following the above described method. The number of samples was similar to those found in previous publications on copy number signatures^{8,9}. K ranged from three in testicular germ cell tumours to eight in five cancers. For 11 cancer types the mathematically optimal set of signatures was also the representative of other solutions. For the other 9 cancer types we picked a representative solution over the mathematical solution. In total we derived 128 cancer-type enriched signatures.

Merging to a signature compendium

As the final step, we merged the pan-cancer and the cancer-type enriched signatures to create a compendium of CIN signatures. We compared signatures by cosine similarity and derived a threshold to merge two signatures using simulation. Based on the set of 138 pan-cancer and cancer-type enriched signatures (138x43 matrix), we simulated 1,000 sets of signatures using a Dirichlet process from MCMCpack¹¹³ (R package, v1.4-7), maintaining

the proportion of 0s and signature components summing to 1 for each signature. Descriptive analyses (heatmaps) were then used to display (i) the distribution of cosine similarities between the signatures, (ii) the simulated signatures, and (iii) the distribution of cosine similarities between the original signatures and the simulated signatures. These analyses showed that the simulated signatures were similar to the original ones. Therefore, we used the quantile 0.999 of the distribution of simulated cosine similarities, which was equal to 0.74, as a cosine similarity threshold.

In order to merge the pan-cancer and cancer-type enriched signatures, we employed a number of strategies to avoid redundancies amongst the signatures:

1. We removed cancer-type enriched signatures which were similar to pan-cancer signatures. To do this, we remove all cancer-type enriched signatures that had a cosine similarity over 0.74 with any pan-cancer signature.
2. We removed cancer-type enriched signatures which were too similar to each other. To do this we selected a representative signature (highest activity for the largest number of samples) from the groups of cancer-type enriched signatures that had a cosine similarity over 0.74.
3. We removed cancer-type enriched signatures, which constitute a combination of pan-cancer signatures. To do this we performed non-negative least squares on each pair of pan-cancer specific signatures and cancer-type enriched signatures. For any combination which showed a reconstruction error below 0.1, the cancer type enriched signature was removed.

The combination of these three strategies reduced the 128 cancer-type enriched signatures to 7. No combination of pan-cancer signatures resulted in a cancer-type enriched signature, and vice-versa, therefore the 10 pan-cancer signatures were combined with the 7 cancer-type enriched signatures to yield a set of 17 signatures of CIN in human cancers. These signatures were then named by their rank determined by the number of samples with non-zero signature activity, ie. their pan-cancer prevalence.

Once defined, the final signature activities for all 6,335 samples with detectable chromosomal instability were computed using the linear combination decomposition function from YAPSA⁸¹ (R package, v1.12.0) on the original input matrix.

Estimating the number of CNAs per signature

We fitted a linear model on the raw signature values to the number of CNAs in each sample. We set the intercept to zero, to only allow positive coefficients. Interaction terms between the signatures were avoided as they resulted in negative coefficients, indicating that the model overfitted. From the model we extracted the coefficients and multiplied with the raw signature values for each sample to get the estimates of CNAs produced by a signature (**Supplementary Fig. 38**).

Assessing feature encoding performance

Simulating CIN

In order to demonstrate the robustness of our approach, we simulated the activity of five types of chromosomal instability described in the literature by introducing copy number changes into the human genome. These included:

- 1) chromosome missegregation via mitotic errors (CHR)¹¹⁴,
- 2) large-scale state transitions via homologous recombination deficiency (LST)¹⁹,
- 3) focal amplification via ecDNA circularisation and amplification (ecDNA)⁶⁵,
- 4) early whole-genome duplication via cytokinesis failure (WGD early)³³, and
- 5) late whole-genome duplication via endoreduplication (WGD late)³³.

We simulated a total of 240 samples. We modelled WGD by setting a tetraploid background before inserting other CNAs (early-WGD) or by multiplying all copy number states after all CNAs were placed (late-WGD) in proportions roughly matching what is expected in cancer genomes³¹. For any given sample, three of the five mutational processes were active. Half of the samples had one dominant signature and the other half had two (**Supplementary Tables 3, 4**). Given that most WGD events are early events, the ratio between early and late WGD samples was 4 to 1 ($N_{\text{early}}=16$ and $N_{\text{late}}=4$ per partition)¹¹⁵.

Supplementary Table 3: Distribution of samples exposed to differently active mutational processes

x - Signature is active with low activity (background)

D - Signature is active with high activity (dominant)

Partition	Large-scale transitions (LST)	ecDNA (ecDNA)	Chromosome missegregation (CHR)	Whole-genome duplication (WGD early + late)	N	Sum
1 dominant signature	D	x	x		20	20
	x	D	x		20	40
	x	x	D		20	60
2 dominant signatures	D	x	D		20	80
	D	D	x		20	100
	x	D	D		20	120
2 dominant signatures plus WGD	D	x		D	16+4	140
	x	D		D	16+4	160
		D	x	D	16+4	180
		x	D	D	16+4	200
	D		x	D	16+4	220
	x		D	D	16+4	240

We used 23 chromosomes (incl. X, excl. Y) of the hg19 reference genome and ignored gender-specific chromosome constellations. The average number of affected chromosomes, number of CNAs produced by a mutational process, as well as the segment size and induced copy number change, were taken from literature where possible (**Supplementary Table 5**).

We could not find evidence in the literature for two parameters for our simulation: the changepoint for the LST/HRD process and the average number of CNAs per chromosome for the focal amplifications. Fortunately, our previous work on large numbers of high-grade serous ovarian cancer genomes (120 published⁸, ~2000 unpublished) allowed us to qualitatively choose these parameters as means of Poisson distributions.

Firstly, the changepoint for the LST process was not mentioned in the original paper¹⁹. In the ovarian genomes we had observed that most genomes coming from BRCA1/2 germline deficient samples tended to have single copy number changes. Therefore, we assumed a single copy loss / gain changepoint, which has also been qualitatively reported in the literature¹⁸.

Secondly, for the average number of CNAs per chromosome produced by focal amplifications, Deshpande et al.¹¹⁶ stated that on average 5 rearrangements were involved per amplicon using deep WGS. However, in the ovarian genomes, where we used shallow WGS, we qualitatively observed that the number of events per amplicon was lower. This is expected given the lower resolution of the technology. The SNP6 array technology used in this study also has lower resolution than deep WGS, thus we reduced the number to 3.

Supplementary Table 4: Mean affected chromosomes by simulated mutational processes

		LST (HRD)	ecDNA	Chrom. misseg.	WGD
Affected chromosomes (mean)	Dominant	8 ¹⁹	4 ¹¹⁶	8 ^{75,76}	23
	Background	2	2	4	-

Supplementary Table 5: Details of the four mutational processes

Full name	ID	Average CNAs per chromosome (Poisson)	Segment size	Changepoint	Background ploidy
Large-scale transitions (LST)	LST	7 ¹⁹	5 - 30MB ¹⁹	0.8 - 1.4 (loss and gain possible)	2
ecDNA	ECDNA	3	1MB - 3MB ¹¹⁷	7-50 ⁶⁵	2
Early whole-genome duplication	WGD (early)	1	Whole chromosome	2	4
Late whole-genome duplication	WGD (late)	1	Whole chromosome	x2 of existing segments	4
Chromosome missegregation	CHR	1	Whole chromosome	0.8 - 1.2 (loss and gain possible)	2

We allowed for multiple mutational processes to be active on the same chromosome by nesting CNAs (smaller CNAs fully overlapped by larger CNAs), but did not allow CNAs to overlap by only one breakpoint.

Homozygous deletions were rare and were only allowed on diploid genomes where an overlap of a HRD-related loss of DNA happened on top of a loss-of-heterozygosity (LOH) produced by a chromosome missegregation. Deletions with a size over 10MB were deemed artificial and were rescued to a subclonal copy number state around 0.5¹⁸.

Activity of mutational processes were defined by randomly assigning them to individual chromosomes. The number of affected chromosomes was drawn from a Poisson distribution with the mean dependent on the strength (dominant or background) and the type of mutational process (**Supplementary Table 4**). We allowed for multiple signatures to target the same chromosome. The number of CNAs per chromosome and mutational process were also drawn by a Poisson (**see Supplementary Table 5 for the means**). If the number of CNAs drawn times the maximum allowed size of the segments was larger than the available space on the chromosome (length minus the centromeric regions) then first, the number was corrected to the maximum number of CNAs that can be accommodated if they all had the maximum size. If this did not resolve the issue, then the maximum allowed segment size was set to 2/3rds of the available space on the chromosome and the number of CNAs set to 1. The latter case could happen e.g. for large-scale transitions which can be up to 30MB but the smallest chromosome is around 50MB. If the number of CNAs drawn was zero, it got rescued to 1.

Then we drew the segment sizes and copy number for each CNA and sorted the CNAs by size. By looping over each CNA, we first determined all regions on the chromosome where this CNA could be placed. The rules were:

- Not overlapping the ends of the chromosomes
- Not ending or beginning in the centromeric region of the chromosome
- Not ending or beginning in other CNAs but beginning and ending in the same CNA (nested CNAs) was allowed

This produced a list of allowed regions of the chromosomes. We sampled uniformly from the regions correcting by their size. This way each base pair position was equally likely to get drawn.

For large-scale state transitions (LST) and chromosome missegregation (CHR), the output copy number changes were around 1 (**Supplementary Table 5**) and we allowed for gains or losses to happen. For chromosome missegregation the probability of a gain was 50%, and for LST it was 25% since losses were stronger associated with HRD in general¹⁸.

The resulting profiles contained a mixture of multiple mutational processes across the genome (**Supplementary Fig. 3**). The profiles were then processed in the same way the TCGA samples were processed (“Deriving signatures”, **Methods page 19**).

Creating gold-standard signatures

As we simulated mutational processes independent of any copy number signatures, we needed to create a gold-standard set of signatures to compare to. To do this, we estimated which feature components should have weight, and which feature components should not have weight, given the characteristics of the simulated copy number (**Supplementary Table 5**). For components which we determined that should receive a weight by the NMF, we assigned a weight of 1 (**Supplementary Fig. 4, shaded bars**).

Identifying simulated signatures

We applied our feature encoding to the simulated genomes then identified signatures using our methodology in the same manner in which it was applied to the TCGA data (“Deriving signatures”, **Methods page 19**). This process yielded 5 signatures, which we then compared to the gold standard (**Supplementary Fig. 4, fully colour bars**). Reassuringly, this showed a good match between the simulated CIN types and the identified signatures: Average cosine similarities was at 0.64 (**Supplementary Table 6**). All permutations tests were significant except for the CHR signature (Permutation test with 10,000 iterations of cosine similarities with FDR-correction). This can be explained by additional small weights on other components, decreasing the similarity value.

Supplementary Table 6: Similarity between gold-standard and identified signatures

Signature	Cosine similarity	q-value
HRD/LST	0.64	0.0005
ecDNA	0.62	0.0005
CHR	0.32	0.46
Early WGD	0.58	0.002
Late WGD	0.58	0.0005
Average	0.64	NA

The signature activities largely agreed with the simulated distributions across samples, with each sample defined by their own set of dominant signatures (**Supplementary Table 3, Supplementary Fig. 5**). We simulated each mutational process to be dominant in a subset of samples (N=100 for chromosomal missegregation, ecDNA, large-scale state transition; N=96 for early whole-genome duplication; N=24 for late whole-genome duplication). The overall accuracy to identify samples with a dominant signature was 78.3% (329/420 dominant samples; **Supplementary Fig. 6**). The overlapping Poisson distributions used to draw the number of affected chromosomes and the number of CNAs adds noise, making perfect accuracy difficult. Given the Poisson parameters we chose, we expected around 20% of samples with background signatures to have levels as high as dominant samples and vice versa. For example, the expected overlap between Poisson distributions with means 8 and 4 is around 20%. These means were used to determine the number of affected chromosomes for dominant and background LSTmutational processes (**Supplementary Table 4**).

In the “Copy number feature encoding” section, **Methods page 13**, we discussed the unique challenges of working with CNAs. One example of this was the identification of chimeric chromosomes, which involve losses or gains of multiple CNAs across the genome. To test that our approach can appropriately detect loss or gain of chimeric chromosomes, we added these events to our simulated genomes. For each simulated sample, we identified an anchor chromosome and a segment from that chromosome that is outside the centromeres. We took its copy number state and identified approximately 7 other target segments from other chromosomes with the same copy number state and which are also outside the centromeric region. The exact number of target segments were drawn from a Poisson distribution with mean 7. For all anchor and target regions, we either uniformly decreased or increased the copy number state by one. We then tested whether any of the signatures representing whole-chromosome change were perturbed. This analysis showed that CX14, a signature of whole chromosome changes, significantly increased upon addition of chimeric chromosomes suggesting our signatures can indeed encode these events correctly (Welch’s t-test with FDR correction; $q\text{-value}=2.54 \times 10^{-06}$).

Single-count versus multi-count feature encoding

Our feature encoding allows a single CNA to contribute probability to multiple features. This is different from other mutational signature encodings where a single mutation only contributes counts to a single feature. Therefore, to validate that this multi-feature encoding approach is more appropriate for encoding signatures of CIN, we compared our approach to a single-count feature encoding inspired by an alternative method for obtaining copy number signatures¹¹⁹. We adapted the approach to our simulated data (“Assessing feature encoding performance”, **Methods page 22**) by classifying each copy number segment into one of 28 categories. These categories were designed to match those found in the preprint as closely as possible, however, as we used total copy number (rather than allele specific) in our simulation, there was not a direct match to the preprint. We used the same segment length and copy number categories, however, genotype was limited to “homdel” when copy number was equal to zero, or “non-homdel” when copy number was non-zero; as opposed to classifying segments as “homdel”, “LOH” and “het” based on the copies of each allele. While this approach is different to that reported in the preprint, we do not expect the results to reflect substantially different signatures to what would be derived using the original method. Given this encoding we used the same approach as the preprint to identify signatures (SigProfilerExtractor, v1.1.4³⁸).

SigProfilerExtractor identified 4 signatures (the true number of signatures was 5; **Supplementary Fig. 43**). We constructed 5 gold-standard signature definitions using this encoding for LSTs, ecDNA, CHR, WGD-early and WGD-late, and compared these to the 4 identified signatures (**Supplementary Fig. 44**). Of the 5 gold-standard signatures, 2 were identified correctly, 1 could be inferred from combinations of the signatures, and 2 were not detected (**Supplementary Table 7**). Signatures A and B represented LSTs on different ploidy backgrounds. Signatures C and D represented ecDNA on different ploidy backgrounds. While there was no signature directly representing WGD, this could be inferred via the different ploidy backgrounds. Signatures for chromosome missegregation events and early-WGD were not detected.

As the single-feature encoding uses absolute copy number, it is expected that the

gold-standard signatures will be artificially split across different ploidy backgrounds. This is precisely why we removed copy number from our multi-feature encoding.

CHR, WGD-early and WGD-late are captured by multi-feature encoding due to the ability to represent many biological properties of CNAs across the feature space. For example, early and late WGD albeit having similar copy number change points, can be separated by the segment size feature (components 16-22 vs 5-22) as well as the breakpoint per 10MB (components 1 vs 2-3) and breakpoints per chromosome arm features (components 1 vs 2-3). In summary, using this set of features, a single feature encoding does not allow the NMF to identify the underlying mutational processes, whereas the multi-feature encoding does.

Supplementary Table 7: Comparisons of different encoding approaches for the discovery of CIN producing mechanisms

	Identified by...	
Mutational process	Multi-feature encoding	Single feature encoding
Correct number of signatures	Yes	No
Early WGD	Yes	No
Late WGD	Yes	Indirectly
Large-scale state transition	Yes	Yes
Chromosome missegregation	Yes	No
ecDNA	Yes	Yes

Robustness analysis

Signature stability assessment via simulation

Signature activity stability

We tested how stable the signature activities were by performing a Monte Carlo simulation where we added noise to the number of CNAs and their characteristics and then measured how strong the signature activities changed.

For each value of the segment length and the copy number changepoint feature, we added or subtracted up to 10% of their value which we drew from a Gaussian distribution with a mean of that value and a standard deviation of one-twentieth of the mean. If a drawn value was larger than 10% we discarded that value and drew again. This procedure had the effect that longer or higher copy number segments were modified stronger than smaller or lower copy number segments.

In addition, we modified the breakpoints per chromosome arm, breakpoints per 10 Mb and length of oscillating copy number chain features by drawing the same number of values from a sample's feature distribution with replacement. In order to avoid a strong deviation from the original sample feature distribution, we applied a maximum change of $\pm 10\%$ to the number of CNAs. If a draw resulted in a higher change in CNAs, then we redrew values until the number of CNAs was below 10% change from the original value.

One run of the simulation included a change of all five feature distributions for 478 samples with a subsequent computation of the signature activities. In total, we ran the simulation 1,000 times. We plotted the interquartile range (IQR) for all samples and all 17 signatures (**Extended Data Fig. 5**).

Signature definition stability

Here, we tested the impact of noise on the signature definitions. Based on the procedure described above, we added noise to all five feature distributions for all 6,335 samples. Then we derived the pan-cancer signatures in the same manner as described above ("Deriving signatures", **Methods page 19**) but using only solutions with K equal to 10 (the number of the original pan-cancer signatures). In total, we ran the simulation 100 times.

We tested how many of the derived signatures matched the original pan-cancer signatures (**Supplementary Fig. 7a**). An adequate match was considered when cosine similarity between original and noisy signature definitions was higher than 0.85. 7 of the 10 pan-cancer signatures had a match across all 100 noisy matrices, with the remaining 3 signatures matching 98 or 99 out of 100 times. We plotted the signature definitions of the original pan-cancer signatures with standard errors for the component weights taken from the 100 simulations (**Supplementary Fig. 7b**).

Signature-specific thresholding

To ensure robustness of the signature activities and to enable trust in small signature activities, we defined signature-specific thresholds. Based on the 1,000 simulations, we used the non-zero simulated values from samples with true zero activity. We fitted a single Gaussian distribution to the non-zero values with the function `Mclust` from `mclust`¹²⁰ (R package, v5.4.6) and took the quantile 0.95 as a threshold (**Extended Data Fig. 5**). To calculate pan-cancer prevalence we counted in how many samples each signature activity was above its specific threshold. This approach gives identical results for pooled data or merging prevalence per cancer type.

Signature stability across genomic technologies

In order to test the robustness of the signatures across sequencing technologies, we used 478 TCGA samples with detectable levels of CIN that were also part of the PCAWG project and derived signature activities and definitions in five settings:

1. SNP 6.0 without matched normal,
2. WGS downsampled to SNP 6.0-positions,
3. WGS downsampled to shallow WGS,
4. WES on-target reads,
5. WES off-target reads.

For each setting, we fitted absolute copy number profiles (step 1), extracted signature activities for the 17 compendium signatures and derived 10 signatures (step 2) and compared them to the 10 pan-cancer signatures (**Extended Data Fig. 6**; step 3). The data set was not big enough for cancer-type specific analyses. For approaches 1, 2 and 4 we used the same ASCAT (R package, v2.4) as described in section “Inferring absolute copy number from SNP 6.0 arrays” (**Method page 10**) available at <https://github.com/VanLoo-lab/ascat>. For the two approaches using binned reads (3 and 5), and for deriving the copy-number profiles of cell lines, we used a new method based on ASCAT, called ASCAT.sc (R package, v0.1, manuscript in preparation) which is currently available here: <https://github.com/VanLoo-lab/ASCAT.sc>.

Step 1: Absolute copy number fitting

Approach 1: SNP 6.0 without matched normals

For many bulk tumour SNP 6.0 profiles, the non-tumour contaminant pulls the B-allele frequency (BAF) of heterozygous SNPs close to 0.5, such that $0 < \text{BAF}_{\text{het}} < 1$ and it can be distinguished from the BAF of homozygous SNPs, i.e. $\text{BAF}_{\text{hom}} = 1$ or $\text{BAF}_{\text{hom}} = 0$. Using this rationale, ASCAT proposes a built-in function to infer which SNPs are heterozygous from the tumour profile only. The parameters of this function are based on the noise profile of the platform and the array design, as manually defined after visual inspection of multiple profiles from the same platform. In this case for Genome-Wide Human SNP Array 6.0, the function expects a range of BAF values for homozygous SNPs ($\text{BAF}_{\text{hom}} < 0.04$ or $\text{BAF}_{\text{hom}} > 0.96$) and a minimum fraction of noisy SNPs ($> 4\%$ expected with $\text{BAF}_{\text{hom}} < \text{BAF}_{\text{noisy}} < \text{BAF}_{\text{het}}$), as well as a platform-specific maximum fraction of heterozygous SNPs (25%) and minimum fraction of homozygous SNPs (67%).

To identify heterozygous SNPs, the BAF of each SNP is first flipped around 0.5 such that all BAF values are within $[0, 0.5]$, i.e. if $\text{BAF} > 0.5$, $\text{BAF}_{\text{flipped}} = 1 - \text{BAF}$. Then the fractions of homozygous f_{hom} , noisy f_{noise} , and heterozygous f_{het} SNPs are determined from the platform-specific parameters and the sample's BAF values: from the total number of SNPs N , the $f_{\text{hom}} \times N$ SNPs with a BAF closest to 0 are labelled homozygous; from the remaining SNPs, $f_{\text{het}} \times N$ SNPs with the lowest distance between their BAF and a local median of the BAF within a window of 301 (of the non-homozygous) SNPs centred around the SNP positions are labelled heterozygous; the remaining SNPs are labelled noisy and left out from downstream analyses.

While this methodology allows to identify heterozygous SNPs in most unbalanced segments, in tumours for which purity is close to 100%, the BAF of heterozygous SNPs in regions that are largely unbalanced, and especially with loss of heterozygosity (LOH), is indistinguishable from the BAF of homozygous SNPs, i.e. $\text{BAF} = 1$ or $\text{BAF} = 0$. Thus, highly imbalanced segments in pure samples are devoid of identifiable heterozygous SNPs. Moreover, because ASCAT's segmentation step only considers heterozygous SNPs, this leaves large blindspots for the identification of copy-number segments in those samples, which would significantly influence our signature framework.

To counter this, we have added a step after the identification of heterozygous SNPs, which rescues SNP positions in segments of $\text{BAF} > 0.75$ and where the density of heterozygous SNPs is much lower than the average in non-LOH regions ($< 20\%$). We rescue as many SNP positions from the SNPs labelled as homozygous, as to reach a 20% density of heterozygous SNPs within the segment. To avoid inducing spurious segments in the BAF of rescued SNPs, we do not rescue SNPs at random but rather the SNPs with BAF values closest to $\text{BAF} = 0.5$. We then run ASCAT on the union of the SNPs identified as heterozygous in the first step and those rescued in segments depleted from heterozygous SNPs.

Approach 2: WGS downsampled to SNP 6.0 positions

From PCAWG WGS (hg19) data, we considered allele counts from alleleCounter (v4.1.0, default parameters with the “--dense-snp” option; <https://github.com/cancerit/alleleCount>) at SNP 6.0 loci. For each sample, loci with at least 1 count in the tumour and 10 counts in the matched normal were used to derive logR and BAF. LogR was then corrected for GC content and replication, following the methodology implemented in the Battenberg R package (<https://github.com/Wedge-lab/battenberg>). ASCAT was then used with different penalty values of interest and a gamma value set to 1.

Approach 3/5: Bin size estimation allowing for downsampling to shallow WGS

As genome binning, or lack thereof, differs between SNP 6.0 array copy number calling and shallow WGS approaches, we sought to identify a genomic bin size which most accurately recapitulates the segment feature distribution seen in SNP 6.0. Segment size was selected as an appropriate measure of the similarity in sample segmentation to compare between differing genomic technologies. A set of 40 ovarian cancer WGS tumour samples from the PCAWG^{54,55} study were used to verify segment size agreement due to a high variability in segment lengths and high segment counts per sample. WGS samples were downsampled to

between 3-15M reads, with consideration for ploidy, purity, bin size (manuscript in preparation).

Segment size distributions were calculated for five genomic bin sizes (5kb, 15kb, 30kb, 50kb, and 100kb) and compared with segment size distributions generated from matched SNP 6.0 array segmentation as described in section “Inferring absolute copy number from SNP 6.0 arrays” (**Methods page 10**). Segment size distributions were tested against the reference set using a Mann-whitney U test after removing segments longer than 10Mb. We determined and selected a bin size of 30kb because its the segment size distribution resulted in a non-significant change (or least significant change) and therefore most accurately replicated the segmentation profiles in matched SNP 6.0 array data (30kb P -value = 0.11, all other bin sizes P -value < 0.001).

We downsampled WGS samples to appropriate read counts, based on the bin size determined prior, sample ploidy, and sample purity and performed absolute copy number fitting as described in section “Shallow WGS: Absolute copy number fitting”.

Approach 3: Shallow WGS: Absolute copy number fitting

To infer copy numbers in shallow coverage WGS, we defined 30kb non-overlapping windows along the genome, excluding centromeres and regions of undefined sequence, i.e. “N”, in the reference genome hg19 (see section above “Approach 3/5: Bin size estimation allowing for downsampling to shallow WGS”).

Then, we empirically identified low-quality bins akin to Scheinin et al.⁵⁰, which can be bins of low mappability or high variability of sequence repeats across individuals. This is done as follows: for 10 normal diploid blood samples from the PCAWG project⁶³ with an estimated fraction of tumour contaminant at 0%, we count the number of reads in each bin and derive the logR across the bins:

$$\text{Log}(R) = \log_2\left(\frac{N_{bin}+1}{\overline{N}_{bin}+1}\right) \quad (1).$$

Where N_{bin} is the number of non-duplicated reads with mapping quality ≥ 30 in the bin, while $\overline{N}_{bin}$ is the average number of such reads across all the bins.

We then corrected the logR for GC content within the bin using the function *loess*¹²¹ (R core package, v.3.6.1), with default parameters. We then ranked the logR values by their counts in each sample and derive a symmetric rank-based score to identify outliers with low or high read counts:

$$s = \log\left(\frac{r}{N+1-r}\right) \quad (2).$$

Where r is the rank of each bin's logR and N is the number of bins. We then summed the scores across samples and took the absolute value. After visual inspection of the profiles, a conservative fifteen percent of the bins with the highest final score were labelled as bad bins and removed from downstream analyses. This was done separately for autosomes and chromosome X.

We then process each tumour sample individually: we counted reads in the remaining 30kb bins and derived the logR as described. The logR was then corrected for the GC fraction of

the bins using *loess*, and segmented using circular binary segmentation (CBS)¹²² using the function *segment*¹²³ (R package DNACopy, v1.64.0) with default parameters. We assessed the influence of the alpha parameter (0.001, 0.01, 0.05), which is the *P*-value threshold used to identify segments with different mean values. After the profile was segmented, we searched the space of combinations of realistic purity and ploidy values (0.01<purity<1, 1.5<average tissue ploidy<5, and tumour ploidy<5) to fit the logR to integer values, solving the logR equation for each segment:

$$\text{Log}(R) = \log\left(\frac{n_T \rho + 2(1-\rho)}{\psi}\right) \quad (3).$$

Where n_T is the total number of copies in the tumour, ρ is the purity and ψ is the average tissue ploidy. We picked the purity and ploidy combination that minimized the distance between n_T and integer values and solved for each segment to obtain a non-rounded total number of copies.

Approach 4: WES: Absolute copy number fitting of on-target reads

From TCGA WES (hg19) data, we considered allele counts from alleleCounter (v4.0.0, default parameters with the “--dense-snp” option; <https://github.com/cancerit/alleleCount>) at SNP loci from the 1,000 Genome Project¹²⁴ (phase 3) falling into exonic regions (GENCODE v19 exome⁷⁸, **Supplementary Table 1**); extended by 1,000bp upstream and downstream). For each sample, loci with at least 1 count in the tumour and 20 counts in the matched normal were used to derive logR and BAF. LogR was then corrected for GC content and replication timing, following the methodology implemented in Battenberg¹²⁵ (R package, v.2.2.9). ASCAT was then used with different penalty values of interest (35, 50, 70, 100, 140) and a gamma value set to 1.

The following rules were used to discard samples:

- No ASCAT solution
- Purity ≥ 0.99 and $1.99 \leq \text{ploidy} \leq 2.01$
- Tumour MAPD ≥ 0.75
- Tumour MAPD ≥ 0.4 and fraction homozygous segments ≥ 0.1
- Differences between logR and BAF segments ≥ 250
- Number of considered SNPs $\leq 75,000$

To pick a single representative sample for multi-sample cases, we considered the following priority: matched blood over matched adjacent normal tissue and most recent analysis date (if available). If multiple samples still fitted these criteria, a sample was randomly selected.

Approach 5: WES: Absolute copy number fitting of off-target reads

We calculated and estimated downsampling read counts using the same methodology and bin size estimates calculated for down sampling of WGS to shallow WGS (see section “Approach 3/5: Bin size estimation allowing for downsampling to shallow WGS”).

To derive the copy-number profiles, we first define an on-target region from the bed definition of the targets in TCGA, adding 1000 base pairs either side of the targets. We then bin the genome into 30kb bins but excluding the on-target regions, i.e. the bins are not necessarily made of contiguous sequences and their genomic width is usually larger than 30kb. We remove bins that overlap with “bad bins” as defined earlier, we derive the logR in each bin,

segment it with CBS, and apply the ASCAT logR-only methodology as described above in section “Approach 3: Shallow WGS: Absolute copy number fitting”.

Step 2: Extract signature activities and definitions

Extraction of the feature values from the copy number profiles were performed as described above in section “Copy number feature encoding” (**Methods page 13**). For identifying the signature activities, we used the linear combination decomposition function from YAPSA⁸¹ (R package, v1.12.0) and the 17 CIN signatures. For identifying new signatures and comparing the definitions to the original 10 pan-cancer signatures, we used the same procedure as described above (“Deriving signatures”, **Methods page 19**) with the notable change that we used only solutions with K=10, the number of the original pan-cancer signatures. As no solution with K=10 were found for the two WES settings and the SNP6-positions only WGS setting, we chose the optimal K (in this case K=7) for the two WES settings and greyed out the figure for the SNP6-positions only WGS setting.

Step 3: Comparison to gold standard (SNP6 with matched normal)

All signature activities were significantly different from a random background distribution of signature activities. For this background distribution, we simulated a 1000 sets of random signature activities for the 478 samples similar to the method described above in section “Merging to a signature compendium”. The expected random cosine similarity was around 0.05 (**Supplementary Fig. 45**, red line). The signature definitions were compared by cosine similarity where we chose a cutoff of 0.95 to define equality. Every one of the 10 original pan-cancer signatures (**Extended Data Fig. 6**, bottom row, y-axis) was found in at least one other setting, showing that the signal can be captured by a small subset of samples from other sequencing settings.

Recommendations for WGS and WES data

We developed our signature approach using TCGA SNP6 data because this represents the largest pan-cancer collection of high-resolution copy number profiled tumours to date. Nevertheless, the resolution of copy number calling from SNP6 data is lower than whole-genome sequence data, which is a limitation of our approach. In the cross-technology platform comparison, we normalised all data to match the resolution of SNP6 arrays and showed that our signature methodology is robust across technology platforms at this resolution (**Extended Data Figure 6**). If the resolution of copy number calling is increased (e.g. by using base-pair resolution whole-genome sequencing derived copy number), the underlying feature distributions change (**Supplementary Fig. 46**). This change can have two interpretations:

1. it may be possible to identify new signatures that were not detected at SNP6 resolution;
2. the SNP6 resolution signatures may not be fit for purpose applied to high resolution data.

For 1), as more pancancer WGS data becomes available, it may be possible to redefine our signatures at higher resolution (if needed).

For 2) we tested this by applying our SNP6 signatures to higher resolution WGS and to WES data processed by multiple different copy number callers (**Supplementary Fig. 47**). As expected, we see a drop in similarity scores for the activities and signature definitions. In some cases (such as ABSOLUTE on WGS), these reductions are minor and it could be possible to use our signatures “out-of-the-box” on these data.

This leads us to the following recommendations when applying the pan-cancer copy number signatures to new data:

- **SNP6 arrays:** Recommended use with and without matched normals using ASCAT with a penalty of 70.
- **WGS:** Recommended use with ASCAT focussing on SNP6 positions or ABSOLUTE on full WGS profiles.
- **Shallow WGS:** Recommended use with ASCAT.sc and a penalty of 0.001
- **WES:** For on-target reads, either ASCAT (penalty 70) or Sequenza can be used. For off-target reads, ASCAT.sc with penalty 0.001 is recommended.

Interpreting signature definitions

The last sections focussed on technical aspects of copy number calling, feature encoding and matrix decomposition to establish the robustness of our approach. The next major step is to interpret the results and establish their biological relevance. Our starting point was the signature definition matrix and we asked: how well do the signature definitions reflect known patterns of CIN? To address this question, we first have to discuss a conceptual difference between a signature and a mutational process.

Signatures versus mutational processes

A mutational process is a faulty biological process that either produces a certain type of genomic change and/or prevents the genomic change from being repaired^{126,127}. The entirety of mutations generated by a mutational process is called a mutational spectrum. Multiple mutational processes can overlap in their mutational spectra. By detecting mutation patterns, we identify signatures or biomarkers of these mutational processes. These signatures might not be defined across all applicable or available features but potentially by a subset only. In these cases, a signature does not fully describe the whole mutational spectrum of a mutational process in all dimensions but acts as a biomarker of its activity. One example is the large-scale transitions (LSTs) in copy number states used to detect HRD via BRCA1/2 inactivation¹⁹. These LSTs are not defined by a specific change in copy number but only by their segment lengths (over 10MB to be predictive). This change in copy number state is not explicitly defined in the original paper¹⁹, nor in the two patents, and is still a valid biomarker that is part of the HRD score^{128,129} which is used in the Myriad MyChoice CDx test and has positive ramifications for thousands of patients annually.

As biomarkers, the signatures do not have to be full descriptions of the mutational spectrum of a mutational process. However, having signatures which are incomplete in their mutational spectra description makes it more difficult to adequately determine their aetiology. For example, in the case of our CX13, it is largely defined by a single high changepoint component. It does not have weight on segment size components, so from the signature it is not possible to get an adequate description of the size of these copy number changes. It is only with a *post hoc* analysis of segments with the defining change point value (at least 10 copies) do we get an estimate for median segment size of 790kB, with an interquartile range between 280kb and 2.7Mb, which is consistent with estimates for ecDNA sizes of 1-3 MB¹¹⁷.

Feature shrinkage is a general property of NMF

We explored where gaps in the signature definitions (like the missing weight on segment size components for CX13) came from and realised that these gaps are often a result of the properties of the NMF. NMF algorithms in general produce sparse results¹³⁰ and the SignatureAnalyzer NMF used in our study is regularised^{38,73,74}. More specifically, we suspect that shrinkage naturally occurs during NMF algorithms whereby stronger components in other signatures are preferred to moderate or weak signals. In order to study this phenomenon, we worked with the SBS signatures which are well-studied single-feature signatures of single base substitutions^{38,131,132}.

We chose three closely related signatures which only differed in the strength of a few components. We randomly indicated if one, two or three SBS signatures were acting on a sample. When one signature was operating in a sample, the total of events simulated ranged from 200 to 300. When multiple signatures were operating in a sample, we assumed that one of the signatures had higher activity levels. Therefore, we modelled that the dominant signature generated a higher number of mutations (from 200 to 300) compared to the other signature(s) operating in the sample (from 50 to 150). In 80% of samples with multiple signatures operating, the dominant signature was the one with higher trinucleotide context variability (SBS10c and SBS5) in order to have an adequate representation of this signature in our simulated dataset. We created 500 samples from those signatures. We then applied NMF and compared the resulting signature definitions with the original definitions. We ran this procedure 250 times to establish a reliable variance.

We did this simulation for two sets of SBS signatures (SBS10a, SBS10b, SBS10c and SBS5, SBS16, SBS92) and for two of the most common NMF algorithms (NMF with Brunet algorithm, R package NMF¹³³, v0.21.0; SigProfilerExtractor, v1.1.4³⁸).

In the first example we tested SBS10a, SBS10c and SBS10d (**Supplementary Fig. 48a**) weights of T[C>A]A and T[C>A]T significantly decreased in the SBS10c signature. Specifically, the weight of T[C>A]A mutations dropped by an order of magnitude from 0.06 to 0.0071 [9.67e-07, 0.031]; while the weight of T[C>A]T decreased from 0.19 to 0.059 [0.0035, 0.13]. In the second example (**Supplementary Fig. 48b**), we combined SBS5, SBS16 and SBS92. In this case, the weight of A[T>C]A significantly decreased from 0.046 to 8.83e-06 [1.76e-12, 1.86e-04] in SBS5.

To confirm that shrinkage is a common property of NMF, we also decomposed signatures from 10 simulations using SigProfilerExtractor which was used for the PCAWG studies. Results from both sets of signature combinations are consistent with those obtained with the Brunet algorithm (**Supplementary Fig. 49**).

In summary, we established that shrinkage of certain features is a common phenomenon when using NMF. It also establishes that the missing weights are most certainly weak or moderate signals. While our experiments were on SBS signatures, we used the same general NMF approach for copy number signatures and assumed that NMF-inherent feature shrinkage explains why not all copy number signatures are defined across all five features (**Supplementary Fig. 8**).

The signature interpretation matrix

To improve the interpretation of our signatures and counteract feature shrinkage, we developed a procedure to generate a “signature interpretation matrix” allowing us to better study the copy number pattern associated with each mutational process. This signature interpretation matrix uses the activity and input matrix to “fill the gaps” and ensure that each signature has weights on each feature.

In detail, we multiplied the sum-of-posterior input matrix by the activity matrix, resulting in a signature by component matrix representing the activity-corrected sum-of-posteriors for each signature. We normalised the matrix to sum to 1, then added it to the original signature

definition matrix, also normalised to sum to 1, to generate the unfiltered signature interpretation matrix (**Supplementary Fig. 9**).

We then filtered the interpretation matrix to only display non-zero values where there was a positive, significant correlation between signature activity and component sum-of-posterior probability. This was determined by first selecting samples in which the signature activity levels were greater than 0, then performing a linear correlation between signature activity levels and the sum-of-posteriors of each component. In the interpretation matrix, we only reported components positively correlated with activity levels ($\beta > 0$ and Benjamini–Hochberg corrected P -value < 0.05). Due to the limited number of samples with CX17 activity, we did not apply the significant threshold to results obtained for this signature. Signature interpretation values were then normalised per signature and feature for plotting purposes (**Extended Fig. 4**).

Validation of the signature interpretation matrix

We validated the signature interpretation matrix procedure using our simulated signatures. We first calculated signature-specific thresholds for the simulated signatures. To do this, we used the non-zero simulated values from samples with true zero activity, and took the quantile 0.80 as a threshold to shrink the values below this to zero. We then calculated a signature interpretation matrix for the 5 signatures identified from the simulated genomes, and compared this to the gold-standard signatures (**Supplementary Fig. 50**). This showed an increase in the average cosine similarity compared to the definition matrix from 0.64 to 0.75 (**Supplementary Tables 6, 8**). In addition, the standard deviation reduced down from 0.13 to 0.09, suggesting more consolidated results and all cosine similarities were significant (Permutation test with 10,000 iterations of cosine similarities with FDR-correction).

Supplementary Table 8: Similarity between gold-standard signatures and interpretation values of identified signatures

Signature	Cosine similarity	q-value
HRD/LST	0.75	0.0005
ecDNA	0.57	0.01
CHR	0.71	0.0005
Early WGD	0.71	0.0005
Late WGD	0.54	0.0025
Average	0.75	NA

These results support the use of a signature interpretation approach. The interpretation matrix also demonstrates results consistent with the literature. For example, CX13 shows weight on the segment size feature components 4-6 (300kb to 2MB) which overlaps strongly with literature estimates for ecDNA^{65,134}. Its other features (bp10MB and bpchrarm) suggest highly clustered mutations, in line with biological expectations of this process. In summary, the use of the signature interpretation matrix allows us to better understand the mutational spectra caused by the mutational processes which are captured by the copy number signatures.

Identification of putative signature aetiologies

Our goal was to associate the 17 CIN signatures with putative causes producing CNAs. We derived an initial hypothesis based on the signature interpretation matrix (**Extended Data Fig. 4**) and the signature definitions (*i.e.* the patterns of copy number change encoded by the signature, **Supplementary Fig. 8**). We also used the signature associations with known cancer driver mutations to form the initial hypothesis (**Extended Data Fig. 8a**, **Supplementary Figs. 10-17**). Our explicit assumption was that the driver gene associations acted as markers for putative pathways involved in the aetiologies and assumed a common pathway deregulation for samples without a driver gene mutation. Below, we describe how we compiled the list of genes of interest, then how we identified the signature aetiologies and then how we associated the signatures with orthogonal data.

Linking signature activities to mutated genes

Selection of 1,146 genes of interest

We performed mutation analysis on 1,146 putative CIN related genes (**Supplementary Fig. 51**) compiled from known pan-cancer driver genes⁷¹, and genes reported as involved in DNA repair, DNA replication, the cell cycle, and chromatin organisation by the Reactome database (Date of access: 03/03/20; ^{72,135}).

The Reactome query resulted in a list of 1,330 proteins known to be involved in DNA repair, DNA replication, the cell cycle, and chromatin organisation. Reactome supplied three types of identifiers: UniProt, ensembl gene identifier, and EMBL identifier. Using the three identifiers, we mapped the proteins to 1,443 ensembl gene entries (ensembl96 on a GRCh37 reference genome). Ignoring the UniProt identifier and removing duplicated entries, the total went down to 1,036 gene names with 1,136 ensembl identifiers.

We added to this list 200 genes reported by Bailey et al.⁷¹ as pan-cancer drivers. We transferred the names to ensembl96 on a GRCh37 reference genome and merged them with the Reactome-derived list of genes. The resulting list had 1,296 ensembl gene identifiers for 1,179 gene names. An additional functional descriptor supplied by Bailey et al. was taken into account for the pan-cancer driver genes: “oncogene” and “tumour-suppressor gene”. If none existed, we assigned the descriptor “ambivalent”. In total, 58 oncogenes, 112 tumour-suppressors, and 30 ambivalent genes were labelled as such.

For all 1,179 genes, we identified their transcriptional start sites (TSS) on a GRCh37 reference genome. Since most genes have multiple TSS, the canonical or strongest proven TSS was obtained by the following decision tree: 1) If only one TSS was present for a gene, then take this one independent of quality; 2) if a GENCODE basic flag was present, take only these, 3) if multiple TSS were present, choose one at random. This decision tree worked for all genes, resulting in 1,179 TSS. After filtering for TSS sites on canonical chromosomes (1-22, X, Y), 1,146 genes were present containing 57 oncogenes, 108 tumour-suppressors, and 30 ambivalent genes.

We used somatic, exonic single nucleotide mutation (SNVs) calls from the TCGA⁶⁹. Gene names, if necessary, were replaced by ensembl96^{136,137} gene names using biomaRt^{138,139} (R package, v2.42.1) to reach consensus names across multiple data sources. The only gene from our list of 1,146 key genes that was not present was *TERC*, which is a long non-coding RNAs. We filtered the set of SNV mutations for synonymous and ambivalent SNVs defined by having received the following variant classification: Silent, Intron, 3'Flank, 3'UTR, 5'Flank, 5'UTR. In addition, we only kept SNVs that passed the TCGA filter criteria. Samples with known hypermutation status have been removed in accordance with previous publications⁷¹.

The set of SNVs did not include germline mutations which had been removed by the NCI Genomics Data Commons (GDC) because of potential issues around patient identification. Germline mutations for *BRCA1* and *BRCA2*, as well as promoter methylations for *BRCA1* and *RAD51C*, were obtained from other TCGA publications (see next section).

For evaluating the copy number state of the 1,146 key genes, we used the same copy number profiles as we did for deriving the CIN signatures. The copy number at the transcriptional start site (TSS) was taken as a surrogate copy number for the whole gene. For each TSS in every sample, we evaluated whether a loss-of-heterozygosity (LOH), and a deletion (DEL) or amplification (AMP) was present. LOH was defined as a state where the b-allele (the allele with the minor copy number) had a copy number of less than 0.4. For deletions the same threshold of 0.4 was applied. Amplifications were called in cases with copy numbers over 5 copies.

Both deletions and amplifications were checked with RNA expression data which were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) website⁵³. Again names of the genes were updated to ensembl96 as described above. Only if an overexpression in the case of oncogenes or underexpression in all other cases matched with an amplification, respectively deletion, was this status kept and moved forward in the analysis.

BRCA1, *BRCA2* and *RAD51C* mutation or hypermethylation

We obtained *BRCA1/2* mutation status from two studies: Maxwell and colleagues⁶¹ identified 37 breast cancer and 63 ovarian cancer germline mutant samples. Wang and colleagues⁶⁰ identified 99 ovarian cancer mutant samples of which 41 had germline and 58 somatic alterations. Given both studies, we had data for 161 patients. Of those, 149 had high-quality copy number arrays and detectable levels of CIN and were therefore included in this study.

We obtained *BRCA1* and *RAD51C* promoter hypermethylation status from TCGA Research Network²⁰. They measured methylation at CpG dinucleotides by using an Illumina GoldenGate methylation assay following previously published protocols¹⁴⁰. In total, they identified 69 ovarian cancer samples, of which 21 were not identified by the other studies.

We used the following order to categorise the mutation status of TCGA samples:

1. Literature sources
 - a. Germline and somatic *BRCA1/2* carriers (OV and *BRCA*)
 - i. Maxwell et al.⁶¹
 - ii. Wang et al.⁶⁰

- b. RAD51C Promoter hypermethylation (OV only)¹⁴¹
- c. BRCA1 Promoter hypermethylation (OV only)¹⁴¹
2. High-impact somatic SNVs (all cancer types, see section above for more details)¹⁴²
3. Loss of heterozygosity (LOH) status (all cancer types)⁵²

This led to the following distribution of mutation status in ovarian (OV) and breast cancer (BRCA, **Supplementary Table 9**).

Supplementary Table 9: *BRCA1*, *BRCA2* and *RAD51C* mutation status in two TCGA cancer cohorts

Mutation group	Number of OV samples	Sum of samples	Number of BRCA samples	Sum of samples
Wild type (WT)	17	17	268	268
WT + LOH at BRCA1/2	382	382	410	410
Germline BRCA2	4	65	5	29
Germline BRCA2 + LOH	23		8	
Germline BRCA1	2		3	
Germline BRCA1 + LOH	36		13	
Somatic BRCA2	0	60	2	23
Somatic BRCA2 + LOH	6		6	
Somatic BRCA1	0		4	
Somatic BRCA1 + LOH	54		11	
BRCA1 promoter hypermeth.	10	21	NA	NA
RAD51C promoter hypermeth.	11		NA	

In total, 146 of 545 ovarian cancer samples (26.8%) harboured either a somatic, germline or promoter alteration of *BRCA1*, *BRCA2* or *RAD51* loci. Another 382 samples harboured LOH at either *BRCA1* or *BRCA2* loci. Only 17 samples had neither a LOH nor any type of mutation at the loci.

Testing for differences in signature activity

Samples were classified as mutant using the following criteria: for oncogenes, only amplifications and point mutations that did not result in a frame shift were considered mutant; for tumour-suppressor genes, we used deletions and all types of point mutations. Amplifications and deletions were only used when they were corroborated by up- or downregulated gene expression.

For *BRCA1* we added two extra comparisons where we evaluated whether a potential mutation was rescued or not. Rescue was defined by a mutation in either *TB53BP1*, *RIF1*, or *MAD2L2*^{143,144}. Not rescued was defined by having no mutation in any of these three genes. As we had prior knowledge about germline mutation and promoter methylation for *BRCA1*, *BRCA2*, and *RAD51C*, we treated these samples as mutated.

To test for differences in signature activity between samples with driver mutations and without a mutation in one of the 1,146 key genes, we scaled and centred the activities per signature (Z-scores). We split the samples into two groups, one group with a mutation and one without. Then we used Welch's t-test to perform a test of equality of means between the two groups. We corrected the resulting 19,006 *P*-values for the false discovery rate¹⁴⁵. Using an alpha of 0.05, we identified 788 genes with at least one significant correlation with the 17 signatures. Of those, 74 genes showed positive associations generating 182 significant correlations with the 17 signatures. We only regarded positive associations in order to avoid compositional data effects (**Supplementary Figs. 10-17, 51**).

Selecting high-confidence gene associations

To ensure that we only regard genes with high confidence when proposing a putative cause, we further filtered the 74 genes down to 32 by applying two additional criteria:

- q-value < 0.005 and
- effect size > +0.4

The following 32 genes were identified as high-confidence genes: *AKT1*, *BRCA1*, *BRCA2*, *CCND1*, *CDK4*, *CDKN2C*, *CIC*, *CUL1*, *ERBB2*, *ERBB3*, *ERCC2*, *FBXW7*, *H3F3A*, *IDH1*, *IDH2*, *MAPK1*, *MYC*, *NFE2L2*, *RAC1*, *RPL22*, *PBRM1*, *PCBP1*, *PIK3R2*, *PMS1*, *PPP2R1A*, *PSMA4*, *SPOP*, *SOS1*, *SOX9*, *TP53*, *U2AF1*, *VHL*. We studied the literature around those genes and collated information on how they are known to influence genomic stability (**Supplementary Table 55**).

Correcting for extrachromosomal DNA amplicons

For the four signatures (CX8, CX9, CX11, CX13) that showed a pattern of amplification (changepoint above 3 copies, **Extended Data Fig. 4**), we studied their relationship with extrachromosomal DNA amplicons (see **ecDNA** section below for how association was determined). We hypothesised that genes amplified via ecDNA amplicons could be a consequence of the mutational process producing those amplicons and can therefore be not causally related to the mutational process. As oncogenes are prone to be amplified by extrachromosomal DNA amplicons^{65,146}, we tested for a shift in signature activities for high-confidence genes amplified in ecDNA amplicons. If found significant, then we assumed the correlation between signature activity and mutation not to be causal and we did not regard this gene for identifying a putative cause. We used a previous study⁶⁵ to identify whether high-confidence genes were part of ecDNA amplicons. For each of the high-confidence genes of the three amplification signatures, using a Welch's t-test we tested whether samples with amplicons for that respective gene had a significantly higher signature activity than samples without amplicons. This led to an additional filtering step, resulting in the removal of association of seven genes with the three amplification signatures. For CX8

we found that amplicons including the genes SPOP, MYC, RAC1 lead to a significantly higher signature activity. For CX9 we excluded CDK4, CCND1, ERBB2 and for CDK11 and CX13 we excluded CDK4.

Correcting for cancer type

The above approach for testing associations between genes of interest and signature activities may be confounded by cancer type (see “Linking signature activities to mutated genes”, **Methods page 38**). One example is VHL mutation, which is nearly exclusively limited to renal cancer. However, having a single tumour type enriched for a specific gene mutation and type of chromosomal instability does not mean that the association is not putatively causal (there is good support in the literature for VHL’s role in regulating mitosis^{147,148}). Therefore, when performing this type of analysis, we faced a conundrum: an analysis without correction potentially reports confounded results, and an analysis with correction potentially excludes strong and real biology (e.g. BRCA2 is mutated in only a small number of tumour types and therefore may not pass correction).

To deal with this, we opted to combine both approaches. We used an uncorrected approach to propose putative aetiologies, then used a corrected approach to see if these aetiologies continued to have support. The corrected approach tested the association between activity and mutations with a multivariate regression model correcting for cancer type.

The original list of genes associated with signatures can be found in Supplementary Table 51 and the corrected genelist in Supplementary Table 53. Nearly all genes in the uncorrected analysis (n=24/32) appeared in the corrected list. The following gene/signature associations appeared in the original list, but did not appear in the corrected list:

- CX1: IDH1, CDKN2C, SOX9, PBRM1, VHL
- CX3: NFE2L2, PIK3R2, PPP2R1A
- CX10: PPP2R1A

These associations have the potential to be confounded and therefore aetiologies based on these need to be treated with caution. Of these, the only ones that pose significant concern are VHL and PBRM1, as these are used to link CX1 with mitotic errors. However, this concern is mitigated by the fact that in the corrected list, tumours with inactivating mutations in CDH1, a key regulator of mitosis, are significantly higher in CX1 activity, supporting the putative aetiology of mitotic errors^{149,150}. These results demonstrate that our proposed aetiologies, which use driver gene mutation associations, are robust to potential confounding effects of tumour type.

Proposing an aetiology and assigning a star rating

The two pieces of evidence we used for proposing aetiologies (ie. putative causes and mechanisms) underpinning the signatures were the high-confidence genes and the pattern of change (signature definition). In order to show our confidence in the evidence supporting a putative cause, we devised a simple star rating from zero to three stars (**Extended Data Fig. 7**).

If high-confidence genes linked a signature to a known type of CIN, we assigned two stars. If there were additionally orthogonal data (i.e. SBS signature or tandem duplication association) supporting the putative cause, we added a third star.

Three stars were given to eight signatures (CX1-CX6, CX8, CX11). Two stars were given to three signatures (CX10, CX14, CX17) with at least one high-confidence gene linked to a known type of CIN but no additional orthogonal data were available. One star was given to three signatures (CX9, CX13, CX16) where there were no high-confidence genes available but some additional orthogonal associations allowed us to propose a speculative aetiology. Zero stars were given to three signatures (CX7, CX12, CX15) where no gene and no orthogonal data were associated with the signature.

Focussed analysis of *CCNE1* and *CDK12*

There is ample evidence for the importance of *CCNE1* and *CDK12* mutations in cancer and chromosomal instability^{151,152}. *CDK12* showed strong associations with CIN signatures CX2 and CX5 but the strong *P*-value correction (for 19,000 multiple tests) scaled their *q*-values to 0.06, above our *q*-value threshold of 0.005. However, those trends support the notion that CX2 and CX5 are associated with class 2 and 3 tandem duplications (**Extended Data Fig. 8, Supplementary Fig. 22**).

CCNE1 was not identified as a driver gene in the TCGA landmark study⁷¹, and therefore did not form part of the above gene enrichment analysis. We therefore performed a separate analysis and found that, classified as pan-cancer oncogene, *CCNE1* is significantly associated with 8 signatures (CX2, CX3, CX4, CX5, CX8, CX9, CX10, **Supplementary Table 10**).

Supplementary Table 10: Testing for differences in signature activity in *CCNE1* mutants and non-mutants Signatures with an effect size of >0.4 and an adjusted *P*-value of <0.005 were bolded.

Signature	# Mutants	# Non-mutants	Mean Mutants	Mean non-mut.	Diff.	Adj. p-val.
CX1	222	5003	-0.8948	0.02564	-0.9204	3.31E-34
CX2	222	5003	0.5389	-0.02311	0.562	8.38E-06
CX3	222	5003	0.7946	-0.03026	0.8249	2.29E-29
CX4	222	5003	0.775	0.01175	0.7632	3.13E-08
CX5	222	5003	0.8564	0.003766	0.8526	9.38E-16
CX6	222	5003	-0.1438	0.04042	-0.1842	0.03818669
CX7	222	5003	0.08714	0.01124	0.0759	0.79330761
CX8	222	5003	0.7787	0.01299	0.7657	4.06E-10
CX9	222	5003	0.7212	0.01558	0.7056	4.54E-10
CX10	222	5003	0.6581	0.005778	0.6523	8.48E-06
CX11	222	5003	0.04352	0.01781	0.02571	0.93504243

CX12	222	5003	-0.01747	-0.01413	-0.00334	0.99471727
CX13	222	5003	0.2693	0.02006	0.2492	0.00036043
CX14	222	5003	-0.2866	9.25E-05	-0.2867	3.23E-11
CX15	222	5003	-0.2824	0.002981	-0.2854	5.41E-12
CX16	222	5003	0.04237	-0.01184	0.05421	0.45347171
CX17	222	5003	0.3179	0.008652	0.3092	0.01508555

Supporting putative aetiologies with orthogonal data

In order to corroborate the putative causes we used an array of orthogonal data:

- two additional patient cohorts and their clinical metadata, (~1,900 patients from the PCAWG project, ~400 patients from the ICGC project);
- 5 types of mutational signatures, (SBS, indel, DBS, ovarian copy number, rearrangement);
- 14 molecular features (somatic point mutations, gene expression, cell cycle score, aneuploidy score, whole-chromosome CNAs, SCNA levels, tandem duplications, loss of heterozygosity (LOH), chromothripsis, kataegis, whole-genome duplication status, telomere length and elongation machinery activity, ecDNA, centrosome amplifications); and
- 11 DNA repair specific features (germline *BRCA1* and *BRCA2* mutations, *BRCA1* and *RAD51C* hypermethylation data, HRDetect response, Myriad myChoice score, *TP53* inactivation score, telomeric imbalances score, large-scale state transition score, LOH score, DNA repair proficiency score, protein expression score for 23 DNA-damage repair genes, PCAWG structural variant with associated microhomologies).

For the sources of the data, please refer to Supplementary Table 1. On the following pages we will give a quick review on where and how we processed these data. The following table contains an overview of the figures where the results are displayed.

Supplementary Table 11: Supplementary figures with the results of orthogonal data analyses

Data source	Suppl. Fig.	Data source	Suppl. Fig.
Chromothripsis	18	Cell cycle score	30
Kataegis	18	Centrosome	24
Loss of heterozygosity	18	Seven HRD metrics	Extended Data Fig. 9b
SBS, DBS, ID signatures	21	Microhomologies of PCAWG SVs	26, 29
Patient age	21 (indirect with SBS1)	Ovarian copy number signatures	25
Whole-chromosome CNAs	18	Gene expression of DNA repair genes	28
Whole-genome duplication	18	Mitotic catastrophe	27
Signatures of ploidy change	18, 52	Impaired HR signatures on ovarian and breast cancers	53
Telomere length and elongation	19, 20	ecDNA	31
Tandem duplication	22		

The results of the analyses on chromothripsis, kataegis, loss of heterozygosity, whole-chromosome CNAs and whole-genome duplication are displayed in **Supplementary Fig. 18**. This figure also contains covariate-specific thresholds allowing for a summary in **Extended Data Fig. 8b**. The cut-off was chosen for each covariate based on how the results clustered in terms of significance and effect-size. Each set of results displayed two clusters - those which had strong significance and large effect-size, and those that had weak or no significance, and low effect-size. For display purposes we highlighted those results that (1) were significant and (2) had large effect sizes. We noticed that some of the significant results clustered with the small effect sizes and de-emphasized them. The cutoffs were different for different covariates because the distributions varied widely.

Chromothripsis

The detection of chromothripsis events was performed for the PCAWG samples⁶³. We only used events that received "Chromothripsis" as a final call. We counted the number of chromothriptic events for the 410 samples with signature activities and tested the Spearman correlation to the signature activities of the 17 CIN signatures. Multiple testing correction was performed according to Benjamini and Hochberg.

The link between HRD and chromothripsis is currently being studied^{23,153}. In our data, we find that IHR signatures CX2 and CX5 are correlated with chromothripsis but not CX3. This may

indicate a mutual exclusivity between chromothripsis and HRD, however, further investigation is required to understand the complex relationships between CIN types.

Kataegis

The detection of kataegis events was performed for the PCAWG samples⁶³. For 498 samples we had with signature activities. We reduced the list of detected kataegis events to those with a significant adjusted *P*-value. We counted the number of kataegis events per sample and tested the Spearman correlation to the activities of the 17 signatures of CIN. Multiple testing correction was performed according to Benjamini and Hochberg.

Loss of heterozygosity (LOH)

Loss of heterozygosity (LOH) segments were obtained from our copy number profiles. LOH was defined as a state where the b-allele (the allele with the minor copy number) had a copy number of less than 0.4. We calculated the proportion of the genome affected by LOH by using the hg19 chromosome sizes. For each signature, we tested the Spearman correlation between this proportion of LOH and signature activity. All *P*-values have been corrected for multiple testing using the Benjamini–Hochberg method.

SBS, DBS and ID signatures

Signature exposures to single base substitutions (SBS), doublet base substitutions (DBS) and small insertion and deletion (ID) signatures were created in Alexandrov et al.³⁸ and were downloaded from Synapse.org. We used signature exposures for both Signature Profiler (SP) and Signature Analyzer (SA). Spearman correlations between SBS, DBS and ID signatures, and the CIN signature activities were performed for the 1900 PCAWG samples. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method for each type of signature (SBS, DBS, ID) and software (SP, SA).

Rearrangement (SV) signatures

The 17 CIN signatures were associated with rearrangement (structural variants, SV) signatures quantified on the ICGC breast cancer cohort²¹ by calculating the Spearman correlation between the signature activities and correcting resulting *P*-values for multiple testing using the Benjamini-Hochberg method.

Patient age

CX1 correlated positively with the “clock-like” SBS1 signature (**Extended Data Fig. 8, Supplementary Fig. 21**) suggesting these errors might also be mediated via a natural ageing process. Therefore, we tested the correlation between CX1 and the age of patients by using a linear as well as a robust linear model correcting for cancer type (R package MASS¹⁵⁴, v7.3_51.5; R core package, v.3.6.1). Pan-cancer, we found no significant relationship (linear model: *P*-value for CX1 and age 0.81; robust linear model: *P*-value 0.88). This result is in line with previous publications which found very few CNAs genomewide in normal tissue, suggesting negative selection¹⁵⁵.

However, we performed the same analysis for each cancer type according to the method of Alexandrov et al.³⁸ and found that for ovarian and lung squamous cell cancer, there was a

positive correlation between age and number of estimated CNAs for CX1 (OV: F-statistic=5.9, q-value=0.048; LUSC: F-statistic=6.8, q-value=0.044). Given that ovarian cancer has one of the longest latencies (over 20 years⁵⁴), it might be that extreme latency periods lead to an accumulation of mitotic errors.

Whole-chromosome CNAs

The smallest canonical human chromosome in the hg19 reference genome is chromosome 22 with 48 Mb. However, SNP 6.0 arrays do not fully cover the whole range of chromosomes. The resulting actual coverage of chromosomes shrunk to 35.2 Mb for chromosome 22 and 37.4 for chromosome 21. We counted a segment as a whole chromosome CNAs if it was larger than 95% of the whole chromosome length covered by SNP6 arrays and had a copy number value greater or smaller than 2. For each signature, we tested the Spearman correlation between counts of whole chromosome CNAs and signature activity. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method.

Whole-genome duplication status and genome instability score

The whole-genome duplication (WGD) status of a tumour was determined based on the overall sample ploidy and the fraction of the genome exhibiting loss of heterozygosity as described previously in D'Amico et al.³¹. Samples with a:

$$Ploidy > 3 - \frac{4}{3} * f_{LOH} \quad (4)$$

with f_{LOH} being the fraction of loss of heterozygosity (LOH), were considered as being whole-genome duplicated in the TCGA data set. For example, samples with a ploidy of 3 and no loss of heterozygosity were classified as having a WGD, as well as tumours with a ploidy of 2 but three-quarters of the genome affected by loss of heterozygosity.

To calculate the genome instability (GI) score, we determined the fraction of each tumour genome that was aberrated, accounting for whole-genome duplication status. Thus, in non-WGD samples all autosomal segments with a copy-number not equal to (1,1) were considered, for WGD samples all segments unequal to (2,2). For the X chromosome copy number we additionally accounted for the sex of the patient.

To test whether the whole-genome duplication status associated positively, negatively, or independently with the activity of a signature, we performed two tests for each signature: a proportion test at zero activity for the proportion of WGD samples compared to the proportion of WGD samples across the 6,335 samples, and a Welch's t-test for samples with activities higher than zero. Only if a signature showed both significantly depleted levels at zero and a significant positive shift in non-zero activities, we positively associated that signature with WGD. If a signature had both an enriched proportion of WGD samples at zero and a significant negative shift in non-zero activities, then we negatively associated that signature with WGD. Otherwise, we classified the signature as independent. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method.

Identification of signatures related to ploidy changes

There are generally two ways a tumour cell can change its ploidy status (where ploidy is measured as the average copy number state observed across the genome): 1) whole-genome duplication which increases the ploidy status by two; or 2) an accumulation of copy number gains which increases the overall ploidy status (similarly for accumulation of losses which reduce the ploidy status). Number 1) is represented by CX4 (**Supplementary Fig. 18**). Number 2) is not directly read out by our signatures. Therefore, we tested if any signature might be linked to a general shift in ploidy status.

By using the function *lm*¹²¹ (R core package, v.3.6.1), we performed a multivariate regression analysis correlating signature activity with tumour ploidy and correcting for CX4 activity and WGD status (**Supplementary Fig. 52**). This analysis shows that CX6 and CX8 seem to be linked to an overall increase in ploidy state, whereas CX2 to a decrease. This is consistent with the proposed aetiologies.

Telomere length and elongation

Telomere length data were obtained from Barthel et al.⁷⁰. The authors estimated telomere length with TelSeq¹⁵⁶ for TCGA samples. We removed estimates for samples based on shallow whole-genome sequencing (sWGS) and whole-exome sequencing (WXS) because we could not evaluate how well telomere length estimates were from off-target reads. A supplementary figure in Barthel et al.⁷⁰ shows significant differences between sequencing technologies. We therefore focussed on estimates derived from whole-genome sequenced samples (WGS). The supplied telomere length ratios were converted to lengths in base pairs. Tumour telomere lengths were subtracted by control telomere lengths. Positive values indicate longer telomere lengths in tumours. For each signature, a robust linear regression model (R package MASS¹⁵⁴, v7.3_51.5) was fitted to the telomere lengths and signature activities. An F-test to test whether the slope of the robust linear model was different from zero was performed with sfsmisc¹⁵⁷ (R package, v1.1-7). The resulting 17 *P*-values were corrected for the false discovery rate according to Benjamini and Hochberg (**Supplementary Fig. 19a**).

Telomerase signature scores (TSS), a gene expression score assessing the activity of the telomere elongation machinery, were obtained from Barthel et al.⁷⁰. For each signature, we tested the Spearman correlation between TSS and signature activity. *P*-values have been corrected for multiple testing using the Benjamini–Hochberg method (**Supplementary Fig. 19b**).

Amplification status for two important genes of the telomere elongation machinery, *TERT* and *TERC*, were obtained from Barthel et al.⁷⁰. We categorised the samples into four groups: both wild type, *TERT* amplified only, *TERC* amplified only, both amplified. For each of the three non-wild type groups, we tested a difference in signature activity to the wild type group with a Welch's *t*-test. *P*-values have been corrected for multiple testing using the Benjamini–Hochberg method (**Supplementary Fig. 19c**).

Gene expression values for *TERT* and *TERC*, were obtained from Barthel et al.⁷⁰. For each signature, we tested the Spearman correlation to both *TERT* and *TERC* gene expression.

P-values have been corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 19d-e**).

The expression status of TERRA, a long non-coding RNA transcribed from telomeres, was also obtained from Bartel et al.⁷⁰. Using Welch's t-test for the binary expression status and signature activities for the 17 CIN signatures, we found a significant depletion of TERRA expression for CX6. TERRA is known to inhibit the elongation of telomeres but is also known to limit genomic instability by protecting telomeres hinting at a more complex picture^{158,159}. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 20a**).

In addition to the telomere elongation machinery, cells utilise a second pathway for telomere elongation called alternative telomere lengthening (ATL) with the chromatin remodeler *ATRX* and *DAXX* being key players. A deactivating mutation in those genes is generally associated with an activation of the ATL pathway^{160,161}. The mutation status of *ATRX* and *DAXX* was obtained from Barthel et al.⁷⁰. Welch's t-tests with the signature activities found no significant difference. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 20b**).

Tandem duplications

Number of tandem duplication events, tandem duplication score and Tandem Duplicator Class categorisation were sourced from Menghi et al.¹⁷. For each signature, we tested the Spearman correlation between the number of tandem duplication events, the tandem duplication score and signature activity. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 22a-b**).

The Tandem Duplicator Class split the samples into a non-TDP group and six different TDP classes (1, 2, 3, 1/3, 1/2, 2/3). The activities of the non-TDP group were compared to the activities of each of the TDP classes using Welch's t-test. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 22c**).

Cell cycle score

The Cell cycle scores (CCS), an additive gene expression score implying the speed of cancer cells going through the cell cycle, for 5466 samples were obtained from Lundberg et al.⁶⁸. We used the same 3 categories as the authors: low, medium (intermediary), high. The signature activities of the high CCS group were compared to the activities of the low and medium group using Welch's t-test. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 30**). We only associated a signature with an increased cell cycle if both the low to medium and the medium to high group comparison were significant.

Centrosome amplification

The CA20 score, an additive gene expression score of 20 genes experimentally proven to be involved with centrosome amplification, for 5437 samples were obtained from Almeida et al.^{22,162}. For each signature, we tested the Spearman correlation between the CA20 score

and signature activity. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 24b**).

We also split the 5437 samples into four groups depending on their dominant impaired HR group. A IHR signature (CX2, CX3, CX5) was dominant in a sample when the scaled z-score was higher than 1.25 and higher than the other IHR signatures. Samples with no dominant IHR signature were put in a “Neither” category. Welch's t-test confirmed an increasing CA20 score from the Neither group to the CX2-dominant, CX5-dominant and CX3-dominant group. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 24a**).

Seven HRD metrics

Sourced from Knijnenburg et al.²⁵, pre-processed by Jordan Griffin (Gerke lab) and published in this Github repository (<https://github.com/GerkeLab/TCGAhrd>), we used the following seven metrics known to correlate with homologous recombination deficiency (HRD): HRD score¹²⁸, *TP53* inactivation score²⁵ (More information in this github repository: <https://github.com/greenelab/pancancer>), telomeric imbalances score¹⁰⁵, large-scale state transition score¹⁹, loss of heterozygosity (LOH) score¹⁸, DNA repair proficiency score (RPS)¹⁶³, protein expression score for 23 DNA-damage repair genes²⁵. For each signature, we tested the Spearman correlation between the seven metrics and signature activity. *P*-values for each metric were corrected for multiple testing using the Benjamini-Hochberg method (**Extended Data Fig. 9b**).

The HRD score is the sum of telomeric imbalances, large-scale state transitions and large loss of heterozygosity events¹²⁸.

Microhomologies of PCAWG structural variants

From the structural variant (SV) data of the PCAWG project⁶³, we extracted the length of microhomologies. Two columns were used: Homlen (length of homology) and homseq (sequence of homology). If the homlen column had a NA entry and the homseq column showed “” [sic], we interpreted this as a length of zero. Structural variants with NAs in both columns were ignored. SVs with 0-1 bp microhomologies were assigned to the nonhomologous end joining (NHEJ) pathway, SVs with 2-20 bp to the theta-mediated end-joining (TMEJ) and SVs with longer stretches of microhomologies were assigned to the single strand annealing (SSA) pathway^{11,164}.

We performed two analyses: One with a focus on samples with a dominant impaired HR signature activity (**Supplementary Fig. 26**) and a broad comparison of signature activity and DNA repair pathway (**Supplementary Fig. 29**).

For the former analysis, we split the 1643 PCAWG samples into four groups depending on their dominant impaired HR group. A IHR signature (CX2, CX3, CX5) was dominant in a sample when the scaled z-score was higher than 1.25 and higher than the other IHR signatures. Samples with no dominant IHR signature were put in a “Neither” category. Welch's t-test confirmed a decrease in NHEJ and an increase from the TMEJ contribution going from the Neither group to the CX2- and CX5-dominant and again to the CX3-dominant

group. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 26a**). For SSA, we identified that the Neither group is driven by outliers. The proportion of samples with zero SSA activity is lower than any of the three IHR signatures but the mean SSA activity is higher than CX2-dominant SSA activity. Therefore, we tested statistical significance by a proportion test of samples having no evidence of SSA activity. Statistical significance between the IHR signatures was tested by a two-sided Welch's t-test with multiple testing correction using Benjamini-Hochberg method. SSA showed a significant decrease in samples with no activity going from Neither to the three IHR signatures. In addition, SSA showed an increase from the CX2- to CX5- and CX3-dominant groups (**Supplementary Fig. 26b**).

For the latter analysis, we split samples into three groups based on their scaled signature activity: depleted (<0), intermediary ($0-1.25$) and dominant (>1.25). For each signature and repair pathway (NHEJ, TMEJ, SSA) Welch's t-tests were performed between the depleted and the dominant sample groups. *P*-values were corrected for multiple testing using the Benjamini-Hochberg method (**Supplementary Fig. 29**).

Ovarian copy number signatures

We compared the 17 CIN signatures to the previously identified 7 ovarian copy number signatures⁸ by computing the cosine similarities of the signature definitions and signature activities called on the 545 TCGA-OV samples (**Supplementary Fig. 25**). We excluded the copy number feature of the ovarian copy number signatures since there was no equivalent feature in the CIN signatures. We mapped the CIN signature feature components onto the ovarian signature components by taking the means of the 43 mixture components and calculating the probabilities of belonging to the mixture components of their respective feature. Using this conversion matrix, we brought the 17x43 CIN signatures to the feature space of the ovarian copy number signatures which has 17x36 dimensions. Each CIN signature was compared to every ovarian copy number signature by using cosine similarity. To test the significance we used a cosine permutation test with 10,000 permutations with the R function `cosinePerm` from `PharmacoGx`¹⁶⁵ (R package, v1.17.1). Taking the maximum from the cosine similarity simulations, we corrected the observed cosine similarity. We decided for this correction to easily visualise highly significant results. For signature activities, cosine similarity, permutation testing, and cosine correction were applied in an identical manner. *P*-values corrected for the false discovery rate according to Benjamini and Hochberg.

Gene expression of DNA repair genes

We explored the activation of DNA repair pathways across 5,186 TCGA samples, which had extracted CIN signature activities and available gene expression data. We initially downloaded the list of genes involved in the different DNA repair pathways from the Reactome database^{72,135} (date of access: 11/05/21), and selected those genes differentially expressed between IHR and non-IHR samples by using a two-sided t-test (*P*-value cutoff of 0.01). IHR samples were those having a high activity of CX2, CX3 or CX5 signatures (threshold of scaled and centered signature activity > 1.25). Then, Spearman correlations were computed between the activity of IHR signatures and the expression of DNA repair genes (Z-scores). All reported *P*-values have been corrected for multiple testing using the Benjamini-Hochberg method.

Ethnicity

Black patients have been shown to experience higher rates of homologous recombination deficiency (HRD)¹⁶⁶. We therefore explored whether the signatures differed in activity across different ethnic backgrounds as annotated by the TCGA. We found that there were substantial differences in signature activities with respect to the patients' ethnicities (**Supplementary Fig. 54**), and that black patients have significantly higher activities in CX3 and CX5, both of which are implicated in impaired HR.

Mitotic catastrophe

Samples with impaired HR signatures CX3 and CX5 experience additional replication stress and impaired DNA damage sensing. The more perturbed the cell cycle checkpoints and DNA repair pathways are, the easier copy number aberrations and structural variants manifest. However, we hypothesised that samples with an abundance of CNAs over SVs might accumulate those during mitosis which is influenced by how quick the cell cycle progresses indicated by replication stress and cell cycle activity and how strong double-strand break (DSB) repair mechanisms are impaired¹⁶⁷.

Using the 1,898 PCAWG samples, we fitted a robust linear model to the number of CNAs and SVs. All samples with weights less than 1 did not fit well in the model and were deemed samples with inferred mitotic catastrophe (**Supplementary Fig. 27a-b**). Indeed, in this group of samples with inferred mitotic catastrophe, high cell cycle scores (CCS) were enriched and significantly higher CX2, CX3 and CX5 contributions with a particular large effect size for CX5 and CX3 (**Supplementary Fig. 27c-d**).

Estimating the number of CNAs produced by each signature, the group of inferred mitotic samples also experienced significantly higher contributions from CX3 and CX5 (**Supplementary Fig. 27e**). When splitting the samples by their cell cycle score (low, medium, high) and comparing these groups to the estimated number of CNAs produced by each of the IHR signatures, a fascinating picture emerges (**Supplementary Fig. 27f-g**). CX2 shows no difference from the background (samples with no significant enrichment in CNAs over SVs). So the increase with cell cycle score is driven by not being able to repair the additional damage accumulating through a faster cell cycle. However, CX5 has constantly and significantly higher contributions to the samples with inferred mitotic catastrophe than the background, probably because a higher rate of replication stress is already present and overloading the repair systems. On the contrary, CX3 shows a negative trend for cell cycle speed. That is, cells are already maxed out in the number of CNAs they can repair due to impaired DNA damages and higher replication stress producing high rates of errors already at low cell cycle speed. Now with increasing cell cycle speed, other signatures (processes) become more active and produce more CNAs, CX3 relative contribution declines as the cells cannot tolerate more CNAs caused by CX3. Therefore, we conclude that most of the excess CNAs produced by CX3 may be caused during the cell cycle potentially by higher replication stress but manifest during mitosis, causing a mitotic catastrophe.

Impaired HR signatures on breast and ovarian cancer samples

Triple-negative breast cancers (TNBC) are known to experience higher rates of homologous recombination deficiency (HRD)^{168,169}. It is therefore of interest how the three impaired HR signatures (IHR, CX2, CX3, CX5) are distributed across TNBC and non-TNBC as well as to ovarian cancer samples, especially for BRCA1/2 and RAD51C wild type samples. We obtained TNBC status from Lehmann et al.⁶⁴ and for each of the three signatures, tested differences in activities between breast (TNBC and non-TNBC) and ovarian (BRCA1/2 mutant and WT) and a control group consisting of other WT BRCA1/2 and RAD51C methylation negative samples in the TCGA cohort (**Supplementary Fig. 53**).

For the purpose of this analysis, we treated all ovarian cancer samples with either a known mutation (germline, somatic or promoter hypermethylation of BRCA1, BRCA2 or RAD51C) or a loss of heterozygosity (LOH) at either BRCA1 or BRCA2 locus as tentatively mutant, leading to a group size of 528 samples. We do not speculate that all 528 samples experience consequences those of a BRCA1/2 loss but given the discoveries of Maxwell et al.⁶¹ and Wang et al.⁶⁰ where they found that LOH is important to convey a phenotype of BRCA1/2 inactivation, we only treated samples with no known mutation as well as no LOH at either locus as true wild type for this analysis (n=17).

Sample sizes for OV WT (n=17) and TNBC (n=7) were small compared to OV mutant (n=528), non-TNBC (n=229) and control (n=2916), which limits the power of any comparison analyses. However, we do find that TNBC samples appear to have activities of CX2 and CX5 matching those of mutant ovarian cancer suggesting these may have a higher propensity to be affected by HRD than non-TNBC, which have activities similar to non-mutant OV and pan-cancer tumours.

ecDNA

We tested for an enrichment of different ecDNA categories⁶⁵ using Fisher's exact test in samples with signature activity versus samples with no signature activity (**Supplementary Fig. 31**). In detail, we used their Supplementary Table 3⁶⁵ containing a list of TCGA sample hierarchically categorised by the presence of amplicons falling into four ecDNA categories (circular → breakage-fusion-bridge → heavily-rearranged → linear → no focal somatic CNA [No-fSCNA]). That means, if a sample had evidence of circular ecDNA, then it fell into the "circular" category independent of how many other categories or amplicons were present. Likewise, only if a sample had no circular ecDNA amplicons, it could have been classified as any other category. We added a fifth category, the presence of any kind of ecDNA which we called "any ecDNA". We did test all categories and displayed the results but given the hierarchical categorisation, we only had the sample size to derive meaningful tests for the circular and any ecDNA category (**Supplementary Fig. 31**). CX13 had by far the strongest association with "any ecDNA" and "circular ecDNA" categories. Given that this signature also had the highest changepoints (>7 and 10 copies), we associated it with ecDNA (**Extended Data Fig. 8b**).

In the CIN signature compendium, the maximum mixture component for the copy number changepoint feature was located at around 10 copies, in stark contrast to the ovarian copy number signature components where the maximum was at 30. This component at 30 copies

change was crucial for defining ovarian signature 6. Given that focal amplifications such as extrachromosomal DNA (ecDNA) are present in a wide range of cancer type but still show cancer-type specific enrichment, particularly in ovarian cancer^{65,170}, we were wondering whether we would miss a signature focussing on extremely high amplifications. We manually added the mixture component around 30 copies to the components of the changepoint feature and rederived pan-cancer signatures. This additional component did not create an additional signature focusing on ultra-high amplification. Instead the additional mixture component was added to the signature already containing the strongest weight on the changepoint component around 10 copies.

Testing for independence between CX2, CX3 and CX5

We present three signatures of impaired homologous recombination (IHR): CX2, CX3 and CX5. Their putative aetiologies were all derived in the same way, accumulating evidence from the signature pattern, gene associations and additional data. In particular, CX2 and CX5, like CX3, are not just associated with mutations in HR genes but also have copy number patterns indicative of defects in HR, which argues that they are specific markers of this process.

To address the question whether CX2 and CX5 are independent forms of IHR, or are their associations driven by co-occurrence with CX3, we performed an association analysis between *BRCA* mutant cases and the activity of the three signatures (CX2, CX3, CX5), using a multivariate regression model which explicitly corrects for the effects of association with the other two signatures, cancer types and *BRCA* mutant cases (**Supplementary Table 12**). The results of this analysis demonstrates that both CX2 and CX5 have significantly higher activities in *BRCA* mutant patients independent of CX3 activity levels and cancer type. This provides evidence that CX2 and CX5 are likely directly associated with *BRCA* mutant status and therefore potentially a result of impaired HR.

Supplementary Table 12: Results of a multivariate regression testing for IHR signatures and *BRCA1/2* mutants

The association between *BRCA1/2* mutation status and signature activity for CX2, CX3 and CX5 was tested while correcting in each analysis for two of the signatures and cancer type.

	CX2 (corrected for CX3 and CX5)		CX5 (corrected for CX2 and CX3)		CX3 (corrected for CX2 and CX5)	
	t	P-value	t	P-value	t	P-value
gBRCA1+LOH	2.2	0.031 *	3.6	0.00002 ***	14.3	<2*10⁻¹⁶ ***
gBRCA2+LOH	-0.3	0.80	-0.2	0.81	10.8	3.0*10⁻⁰⁹ ***
sBRCA1+LOH	2.9	0.00392 **	6.5	8.8*10⁻¹¹ ***	14.2	<2*10⁻¹⁶ ***
sBRCA2+LOH	1.6	0.11	-0.2	0.85	8.7	<2*10⁻¹⁶ ***
BRCA1 hyperm.	1.3	0.19	3.5	0.0004***	6.2	6.6*10⁻¹⁰***
RAD51C hyperm.	1.1	0.28	-1.4	0.15	6.4	1.7*10⁻¹⁰ ***
WT BRCA1/2	= benchmark against which we compared					

Correlating signature activity with drug response

Calculating signature activities for cell lines

Cell line SNP 6.0 array CEL files were downloaded from the CCLE project⁵⁷ and were processed using Affymetrix Power Tools (APT; v2.11.2; standard options). Combined cell line logR and BAF data was split into sample-specific logR and BAF files.

No germline samples were available for the identification of heterozygous SNPs in the cell line data, thus we ran the pipeline for SNP 6.0 without matched normal described above (see section “SNP 6.0 without matched normals”).

In this pipeline, while the rescued SNPs allow for the identification of additional segments from their LogR, by construction their average BAF is never equal to 0 or 1 even in the presence of LOH. This biases purity estimates towards lower values and renders the copy-number fitting inaccurate. To circumvent this problem, we first fitted logR and BAF as per the ASCAT methodology⁷⁹, then searched the space of combinations of purity and ploidy around the same ploidy solutions, but between 95% and 100% purity and only solving for the total copy number using the LogR-only equation for each segment:

$$\text{Log}(R) = \log\left(\frac{n_T \rho + 2(1-\rho)}{\psi}\right) \quad (5).$$

Where n_T is the total number of copies in the tumour, ρ is the purity and ψ is the average tissue ploidy. Akin to ASCAT, we picked the purity and ploidy combination that minimized the distance between n_T and integer values. Only fitting the LogR and not the BAF in cell lines lead to more accurate values of the purity and thus the copy-number profiles.

As fitting of absolute copy number to non-matched cell line data is inherently more challenging than other instances, all cell lines where ASCAT identified a plausible solution were visually inspected to determine fit suitability and excluded cell lines where absolute copy number profiles were incorrect or poorly fitted.

We extracted signature activities for the 17 compendium signatures in 594 cancer cell lines. Extraction of the feature values from the copy number profiles derived from cell lines were performed as described above in section “Copy number feature encoding” (**Methods page 13**). For identifying the signature activities, we used the linear combination decomposition function from YAPSA⁸¹ (R package, v1.12.0) and the 17 signatures of CIN.

CRISPR, RNAi and drug response data integration

Using the depmap portal we obtained essentiality scores of 17,645 genes from CRISPR knock-out screens⁵⁶, essentiality scores of 17,309 genes from RNAi screens⁵⁹, and area under dose response curve (AUC) values for 1,394 drugs across 481 human cancer cell lines from the PRISM screen (HTS002 screen)⁵⁸. After merging these data with cell line copy number profiles, we obtained a total of 382 cell lines with available copy number profiling and gene essentiality data from CRISPR and RNAi screens. From all of these, 297 also had drug response data from the PRISM screen.

Correlation with target gene KO, KD and drug inhibition

Kendall's Tau correlation was performed between CIN signature activities and gene essentiality scores from the genetic perturbation screens. Correlations were filtered for $\tau < -0.07$ and $q\text{-value} < 0.05$ after multiple testing corrections using the Benjamini-Hochberg method. This yielded 7,225 targets significantly associated with a signature from the CRISPR screen and 7,091 targets from the RNAi screen. These targets were then used to test correlation of the signature activity with matched drug inhibition of the target measured via AUC. After filtering (Kendall's $\tau < -0.1$ and Benjamini-Hochberg corrected $q\text{-value} < 0.01$) a total of 44 drug responses were correlated with signature activity via 40 targets.

Conversely, those targets significantly linked to a CIN signature in both CRISPR and RNAi perturbation screens but without matched drug response represented putative novel targets for drug discovery. From an initial list of 319 targets overlapping in both the RNAi and CRISPR based correlation, 297 had unknown targeted therapies. After manually inspecting the target tractability¹⁷¹, 104 of these targets were considered druggable according to their structure or by ligand-based approaches, and after manual literature curation, 49 had a clear prior implication in CIN.

Building a clinical classifier based on signature activities

Preparation of survival data

TCGA Clinical metadata were downloaded from the [TCGA website](#). Clinical metadata for the ICGC breast cancer cohort was from the original publication²¹. Data for PCAWG OV-AU and ESAD-UK were obtained from Synapse⁶³.

Age was converted into years and split into 4 groups: [0, 53), [53, 62), [62, 76), [76+). Stages were simplified by omitting additional stage identifiers after the initial roman letter identifier (e.g. Stage IIA1 became Stage II). Stage IS was treated as NA. We added some missing clinical stages supplied by Liu et al.¹⁷².

Survival analysis

In order to study the relationship between the signature activities, clinical classifiers and survival of patients, we used an univariate (Kaplan-Meier estimates) and a multivariate method (Cox proportional hazard model). Kaplan-Meier estimates (function *survfit*) and Cox proportional hazard models (function *coxph*) were performed using survival¹⁷³ (R package, v3.1-11) and survminer¹⁷⁴ (R package, v0.4.8). For the Kaplan-Meier estimates a logrank test was performed. For the Cox proportional hazard models we included the covariates stage and age category, except for HRDetect classification on the TCGA-OV cohort where we only corrected for age and not stage since there were not enough samples in each stage for the model to converge (n=40).

Clinical classifier based on CX2/CX3 signature activity

From our work on the signature aetiologies, we found that CX2 has no significant relation to predicting platinum-sensitivity in TCGA ovarian cancer patients but CX3 does (**Supplementary Fig. 32**). We devised a clinical classifier exploiting exactly this relation on the 36 germline BRCA1 mutant samples of the TCGA-OV cohort that had an additional loss of heterozygosity at the BRCA1 locus. For each sample, we normalised their 17 signature activities to 1. Then we centred the mean and scaled the variance of the signatures CX2 and CX3 across the 36 samples. This scaling and centering is the first part of the classifier.

The second part of the classifier is checking which signature has a higher value and classify accordingly: if CX2 activity is higher we predicted platinum-resistance and if CX3 activity is higher then we predicted platinum-sensitivity.

Applying this classifier to new samples means that their CX2 and CX3 signature activities are evaluated relatively to germline BRCA1 mutant samples. The scaling worked well for ovarian and breast cancer samples but had to be redone for the oesophageal samples, by scaling the normalised CX3 and CX2 activities only on the PCWAG-ESAD cohort.

We applied the CX2/CX3 classifier to the whole ovarian cohort and found that 27% of the TCGA OV cohort might be platinum sensitive. Extending this classifier pan-cancer, we estimate that at most 8% of all tumours might be sensitive.

Comparison with HRDetect and Myriad myChoice classifier

HRDetect values were taken from Davies et al.⁶ and Degasperi et al.⁶⁶. As in the original publication, the threshold of 0.7 was used to identify positive samples⁶.

Myriad myChoice classifier is based on the HRD score¹²⁸. Positive samples were samples with scores above or equal to 42. HRD scores were taken from Knijnenburg²⁵ et al. and complemented by data from Marquard et al.⁶⁷ for samples with NA values.

Clinical classifier based on three impaired HR signatures

We deployed three additional clinical classifiers taking into account all three impaired HR signatures (CX2, CX3, CX5). Hazard ratios of the three classifiers for the resistant or early death group were at 0.62, 0.57 and 0.67 and therefore inferior to the 0.54 hazard ratio produced by the CX2/CX3 classifier.

Logistic regression on survival outcome

First, we extracted the germline BRCA1 samples from the ovarian TCGA cohort and scaled the activities for the three impaired HR signatures. Then we trained a logistic regression using the function *glm*¹²¹ (R core package, v.3.6.1) on the scaled signatures to predict the outcome of the survival data (death). Then we scaled the signature activities of all TCGA ovarian cancer samples according to the scale and variance from the germline BRCA1 samples and tested the logistic regression classifier by producing a Kaplan-Meier estimate and a Cox proportional hazard model (**Supplementary Fig. 37**, top row).

Support vector machine (svm) on survival outcome

The second classifier was a support vector machine (svm). We trained the model on the same data as the logistic regression and CX2/CX3 classifiers: scaled impaired HR signature activities from the germline BRCA1 samples of the TCGA ovarian cancer cohort. During training to predict survival outcome (death) we avoided overfitting by ten-fold cross validation. Afterwards, the activities of the complete TCGA-OV cohort were scaled according to the germline BRCA1 sample scaling with the svm predicting an early death based (**Supplementary Fig. 37**, middle row).

Support vector machine (svm) on hierarchical clustering

The third classifier was an svm predicting the class membership of two groups produced by hierarchical clustering on the scaled signature activities of the germline BRCA1 samples of the TCGA ovarian cancer cohort. A ten-fold cross validation was used as a regularisation method. The classifier was then applied to the complete TCGA-OV cohort which was again scaled according to the activities of the germline BRCA1 samples (**Supplementary Fig. 37**, bottom row).

Reference genome

The reference genome throughout the whole project was GRCh37 / hg19.

Software versions

If not otherwise stated, the free statistical software R¹²¹ (v3.6.1) was used. If functions outside the standard packages were used, they were mentioned and cited throughout the text. A summary of all non-standard packages can be found below in Supplementary Table 13.

Supplementary Table 13: R packages used throughout the project

Name	Version	Citation	Name	Version	Citation
broom.mixed	0.3.3	¹⁷⁵	lsa	0.73.1	¹⁷⁶
caret	6.0-83	¹⁷⁷	MASS	7.3_51.5	¹⁵⁴
cmprsk	2.2-10	¹⁷⁸	NMF	0.21.0	¹³³
coin	1.4-2	¹⁷⁹	patchwork	1.0.0	¹⁸⁰
compositions	2.0-1	¹⁸¹	PharmacoGx	1.17.1	¹⁶⁵
cowplot	1.0.0	¹⁸²	Polychrome	1.3.1	¹⁸³
data.table	1.12.8	¹⁸⁴	RColorBrewer	1.1-2	¹⁸⁵
DNAcopy	1.64.0	¹²³	reshape2	1.4.3	¹⁸⁶
flexmix	2.3-15	¹⁰⁹	rstatix	0.7.0	¹⁸⁷
ggExtra	0.9	¹⁸⁸	scales	1.0.0	¹⁸⁹
ggplot2	3.1.1	¹⁹⁰	sfsmisc	1.1-7	¹⁵⁷
ggtrastr	0.2.3	¹⁹¹	sjPlot	2.8.9	¹⁹²
ggrepel	0.8.2	¹⁹³	stringr	1.4.0	¹⁹⁴
ggthemes	4.2.0	¹⁹⁵	survival	3.1-11	¹⁷³
ggstance	0.3.3	¹⁹⁶	survminer	0.4.8	¹⁷⁴
ggpubr	0.2.5	¹⁹⁷	tidyverse	1.3.1	¹⁹⁸
gridExtra	2.3	¹⁹⁹	vcfR	1.12.0	²⁰⁰
lemon	0.4.4	²⁰¹	viridisLite	0.3.0	²⁰²
jtools	2.0.3	²⁰³	YAPSA	1.12.0	⁸¹

References

- Bakhroum, S. F. & Cantley, L. C. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell* **174**, 1347–1360 (2018).
- Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* vol. 144 646–674 (2011).
- 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).
- Chakravarti, D., LaBella, K. A. & DePinho, R. A. Telomeres: history, health, and hallmarks of aging. *Cell* **184**, 306–322 (2021).
- Bakhroum, S. F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).
- Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
- Cohen-Sharir, Y. *et al.* Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature* **590**, 486–491 (2021).
- Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
- Steele, C. D. *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell* **35**, 441–456.e8 (2019).
- Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).
- 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).
- Chittaranjan, S. *et al.* Loss of CIC promotes mitotic dysregulation and chromosome segregation defects. *bioRxiv* 533323 (2019) doi:10.1101/533323.
- 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).
- 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).
- Silverman, J. S., Skaar, J. R. & Pagano, M. SCF ubiquitin ligases in the maintenance of genome stability. *Trends Biochem. Sci.* **37**, 66–73 (2012).
- Godinho, S. A. & Pellman, D. Causes and consequences of centrosome abnormalities in cancer. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, (2014).
- 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).
- 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).
- 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).
- Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
- Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- 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).
- Piazza, A. & Heyer, W.-D. Homologous Recombination and the Formation of Complex Genomic Rearrangements. *Trends Cell Biol.* **29**, 135–149 (2019).
- Guirouilh-Barbat, J., Lambert, S., Bertrand, P. & Lopez, B. S. Is homologous recombination really an error-free process? *Frontiers in Genetics* vol. 5 (2014).
- 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).
- Saavedra, H. I., Fukasawa, K., Conn, C. W. & Stambrook, P. J. MAPK mediates RAS-induced chromosome instability. *J. Biol. Chem.* **274**, 38083–38090 (1999).
- 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).
- 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).
- 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).
- 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).
- Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239–2254.e39 (2021).
- Dewhurst, S. M. *et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
- 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).
- Berenjeno, I. M. *et al.* Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. *Nat. Commun.* **8**, 1773 (2017).
- Darp, R., Vittoria, M. A., Ganem, N. J. & Ceol, C. J. Oncogenic BRAF Induces Whole-Genome Doubling Through Suppression of Cytokinesis. *bioRxiv* 2021.04.08.439023 (2021) doi:10.1101/2021.04.08.439023.
- Zhang, Q. *et al.* FBXW7 Facilitates Nonhomologous End-Joining via K63-Linked Polyubiquitylation of XRCC4. *Mol. Cell* **61**, 419–433 (2016).
- 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).
- Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- 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).
- Crockford, A. *et al.* Cyclin D mediates tolerance of genome-doubling in cancers with functional p53. *Ann. Oncol.* **28**, 149–156 (2017).
- 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).
- Goel, S. *et al.* CDK4/6 inhibition triggers anti-tumour immunity. *Nature* **548**, 471–475 (2017).
- Brownlee, P. M., Meisenberg, C. & Downs, J. A. The SWI/SNF chromatin remodelling complex: Its role in maintaining genome stability and preventing tumorigenesis. *DNA Repair* **32**, 127–133 (2015).
- 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. U. S. A.* **101**, 8699–8704 (2004).
- Quinton, R. J. *et al.* Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* (2021) doi:10.1038/s41586-020-03133-3.
- 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. U. S. A.* **106**, 19108–19113 (2009).
- 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).
- Martin, L. P., Hamilton, T. C. & Schilder, R. J. Platinum resistance: the role of DNA repair pathways. *Clin. Cancer Res.* **14**, 1291–1295 (2008).
- Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
- 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).
- Heil, B. J. *et al.* Reproducibility standards for machine learning in the life sciences. *Nat. Methods* **18**, 1132–1135 (2021).

52. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
53. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
54. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
55. Dentre, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *bioRxiv* 312041 (2020) doi:10.1101/312041.
56. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
57. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
58. Corsello, S. M. *et al.* Non-oncology drugs are a source of previously unappreciated anti-cancer activity. *bioRxiv* 730119 (2019) doi:10.1101/730119.
59. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **9**, 4610 (2018).
60. Wang, Y. K. *et al.* Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nat. Genet.* **49**, 856–864 (2017).
61. Maxwell, K. N. *et al.* BRCA locus-specific loss of heterozygosity in germline BRCA1 and BRCA2 carriers. *Nat. Commun.* **8**, 319 (2017).
62. Research, N. *et al.* Cancer Genome Atlas The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
63. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
64. Lehmann, B. D. *et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* **11**, e0157368 (2016).
65. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).
66. Degasper, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).
67. Marquard, A. M. *et al.* Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark Res.* **3**, 9 (2015).
68. Lundberg, A. *et al.* A pan-cancer analysis of the frequency of DNA alterations across cell cycle activity levels. *Oncogene* **39**, 5430–5440 (2020).
69. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).
70. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
71. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).
72. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
73. Tan, V. Y. F. & Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the β -divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).
74. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
75. Weaver, B. A. A. & Cleveland, D. W. Does aneuploidy cause cancer? *Curr. Opin. Cell Biol.* **18**, 658–667 (2006).
76. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
77. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, (2017).
78. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
79. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
80. Affymetrix, Inc. *Frequently Asked Questions: Chromosome Analysis Suite 3*. (ChAS 3.1). https://www.affymetrix.com/support/help/faqs/pdf/chas/chas_3_1_faqs.pdf.
81. Huebschmann, D., Gu, Z. & Schlesner, M. YAPSA: Yet Another Package for Signature Analysis. (2019).
82. Cimini, D. Merotelic kinetochore orientation, aneuploidy, and cancer. *Biochim. Biophys. Acta* **1786**, 32–40 (2008).
83. Lens, S. M. A. & Medema, R. H. Cytokinesis defects and cancer. *Nat. Rev. Cancer* **19**, 32–45 (2019).
84. Mitelman, F., Johansson, B. & Mertens, F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. <https://mitelmandatabase.isb-cgc.org>.
85. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
86. Cisyk, A. L. *et al.* Characterizing the Prevalence of Chromosome Instability in Interval Colorectal Cancer. *Neoplasia* **17**, 306–316 (2015).
87. Duijff, P. H. G., Schultz, N. & Benezra, R. Cancer cells preferentially lose small chromosomes. *Int. J. Cancer* **132**, 2316–2326 (2013).
88. Harbour, J. W. Eye cancer: unique insights into oncogenesis: the Cogan Lecture. *Invest. Ophthalmol. Vis. Sci.* **47**, 1736–1745 (2006).
89. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
90. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
91. Jonsdottir, A. B. *et al.* Tetraploidy in BRCA2 breast tumours. *Eur. J. Cancer* **48**, 305–310 (2012).
92. Pradhan, M. *et al.* Gross genomic alterations and gene expression profiles of high-grade serous carcinoma of the ovary with and without BRCA1 inactivation. *BMC Cancer* **10**, 493 (2010).
93. Chien, J. *et al.* TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. *Nucleic Acids Res.* **43**, 6945–6958 (2015).
94. Ryland, G. L. *et al.* Loss of heterozygosity: What is it good for? *BMC Med. Genomics* **8**, 45 (2015).
95. Nichols, C. A. *et al.* Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nat. Commun.* **11**, 1–14 (2020).
96. Macintyre, G., Ylstra, B. & Brenton, J. D. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends Genet.* **32**, 530–542 (2016).
97. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.* **52**, 1419–1427 (2020).
98. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* vol. 21 53 (2020).
99. Etemadmoghadam, D. *et al.* Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19489–19494 (2013).
100. Murnane, J. P. Telomere dysfunction and chromosome instability. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **730**, 28–36 (2012).
101. Zakov, S. & Bafna, V. Reconstructing breakage fusion bridge architectures using noisy copy numbers. *J. Comput. Biol.* **22**, 577–594 (2015).
102. Malhotra, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* **23**, 762–776 (2013).
103. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer* **1** (2019).
104. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
105. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* **2**, 366–375 (2012).
106. Umbreit, N. T. *et al.* Mechanisms generating cancer genome complexity from a single cell division error. *Science* **368**, (2020).
107. Ng, C. K. Y. *et al.* The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J. Pathol.* **226**, 703–712 (2012).
108. Blei, D. M. & Jordan, M. I. Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**, 121–143 (2006).
109. Grün, B. & Leisch, F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software, Articles* **28**, 1–35 (2008).
110. 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).
111. Devarajan, K. Nonnegative matrix factorization: an analytical

- and interpretive tool in computational biology. *PLoS Comput. Biol.* **4**, e1000029 (2008).
112. Yang, Z. & Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32**, 1–8 (2016).
 113. Martin, A., Quinn, K. & Park, J. H. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software, Articles* **42**, 1–21 (2011).
 114. Jallepalli, P. V. & Lengauer, C. Chromosome segregation and cancer: cutting through the mystery. *Nat. Rev. Cancer* **1**, 109–117 (2001).
 115. Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
 116. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 1–14 (2019).
 117. Liao, Z. *et al.* Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression. *Biochim. Biophys. Acta Rev. Cancer* **1874**, 188392 (2020).
 118. Cheng, J. *et al.* Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat. Commun.* **8**, 1221 (2017).
 119. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *bioRxiv* 2021.04.30.441940 (2021) doi:10.1101/2021.04.30.441940.
 120. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 289–317 (2016).
 121. R Core Team. R: A Language and Environment for Statistical Computing.
 122. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
 123. Seshan, V. E. & Olshen, A. DNACopy: DNA copy number data analysis. (2019).
 124. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 125. Wedge, D., Van Loo, P. & D'Entro, S. Battenberg: Battenberg subclonal copy number caller. (2021).
 126. Volkova, N. V. *et al.* Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.* **11**, 2169 (2020).
 127. Frigola, J., Sabarinathan, R., Gonzalez-Perez, A. & Lopez-Bigas, N. Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. *Nucleic Acids Res.* **49**, 891–901 (2021).
 128. Telli, M. L. *et al.* Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).
 129. Timms, K. M. *et al.* Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Res.* **16**, 475 (2014).
 130. Peharz, R. & Pernkopf, F. Sparse nonnegative matrix factorization with l0-constraints. *Neurocomputing* **80**, 38–46 (2012).
 131. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
 132. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* vol. 149 979–993 (2012).
 133. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367–1471 (2010).
 134. Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* (2021) doi:10.1038/s41586-021-04116-8.
 135. Fabregat, A. *et al.* Reactome graph database: Efficient access to complex pathway data. *PLoS Comput. Biol.* **14**, e1005968 (2018).
 136. Hunt, S. E. *et al.* Ensembl variation resources. *Database* **2018**, (2018).
 137. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
 138. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
 139. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
 140. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
 141. Network, C. G. A. R. & Network, T. C. G. A. R. Integrated genomic analyses of ovarian carcinoma. *Nature* vol. 474 609–615 (2011).
 142. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).
 143. Boersma, V. *et al.* MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5' end resection. *Nature* **521**, 537–540 (2015).
 144. Isono, M. *et al.* BRCA1 Directs the Repair Pathway to Homologous Recombination by Promoting 53BP1 Dephosphorylation. *Cell Rep.* **18**, 520–532 (2017).
 145. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
 146. Kumar, P. *et al.* ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. *Sci Adv* **6**, eaba2489 (2020).
 147. Jonasch, E., Walker, C. L. & Rathmell, W. K. Clear cell renal cell carcinoma ontogeny and mechanisms of lethality. *Nat. Rev. Nephrol.* **17**, 245–261 (2021).
 148. Thoma, C. R. *et al.* VHL loss causes spindle misorientation and chromosome instability. *Nat. Cell Biol.* **11**, 994–1001 (2009).
 149. Li, M. & Zhang, P. The function of APC/CCdh1 in cell cycle and beyond. *Cell Div.* **4**, 2 (2009).
 150. Skaar, J. R. & Pagano, M. Cdh1: a master G0/G1 regulator. *Nature cell biology* vol. 10 755–757 (2008).
 151. Zhao, Z.-M. *et al.* CCNE1 amplification is associated with poor prognosis in patients with triple negative breast cancer. *BMC Cancer* **19**, 96 (2019).
 152. Krajewska, M. *et al.* CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat. Commun.* **10**, 1757 (2019).
 153. Nazaryan-Petersen, L., Bjerregaard, V. A., Nielsen, F. C., Tommerup, N. & Tümer, Z. Chromothripsis and DNA Repair Disorders. *J. Clin. Med. Res.* **9**, (2020).
 154. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. (2002).
 155. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
 156. Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42**, e75 (2014).
 157. Maechler, M. sfsmisc: Utilities from 'Seminar fuer Statistik' ETH Zurich. (2020).
 158. Luke, B. & Lingner, J. TERRA: telomeric repeat-containing RNA. *EMBO J.* **28**, 2503–2510 (2009).
 159. Oliva-Rico, D. & Herrera, L. A. Regulated expression of the lncRNA TERRA and its impact on telomere biology. *Mech. Ageing Dev.* **167**, 16–23 (2017).
 160. Heaphy, C. M. *et al.* Altered telomeres in tumors with ATRX and DAXX mutations. *Science* **333**, 425 (2011).
 161. Dyer, M. A., Qadeer, Z. A., Valle-Garcia, D. & Bernstein, E. ATRX and DAXX: Mechanisms and Mutations. *Cold Spring Harb. Perspect. Med.* **7**, (2017).
 162. Almeida, B. P. D., Vieira, A. F., Paredes, J., Bettencourt-Dias, M. & Barbosa-Morais, N. L. Pan-cancer association of a centrosome amplification gene expression signature with genomic alterations and clinical outcome. *PLoS Computational Biology* vol. 15 e1006832 (2019).
 163. Pitroda, S. P. *et al.* DNA repair pathway gene expression score correlates with repair proficiency and tumor sensitivity to chemotherapy. *Sci. Transl. Med.* **6**, 229ra42 (2014).
 164. Schimmel, J., van Schendel, R., den Dunnen, J. T. & Tijsterman, M. Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. *Trends Genet.* **35**, 632–644 (2019).
 165. Smirnov, P. *et al.* PharmacGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* vol. 32 1244–1246 (2016).
 166. Sinha, S. *et al.* Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. *Nature Cancer* **1**, 112–121 (2020).
 167. Schoonen, P. M. *et al.* Progression through mitosis promotes PARP inhibitor-induced cytotoxicity in homologous recombination-deficient cancer cells. *Nat. Commun.* **8**, 15981 (2017).
 168. Belli, C., Duso, B. A., Ferraro, E. & Curigliano, G. Homologous recombination deficiency in triple negative breast cancer. *Breast* **45**, 15–21 (2019).
 169. Telli, M. L. *et al.* Homologous recombination deficiency and host anti-tumor immunity in triple-negative breast cancer. *Breast Cancer Res. Treat.* **171**, 21–31 (2018).
 170. Gu, X., Yu, J., Chai, P., Ge, S. & Fan, X. Novel insights into extrachromosomal DNA: redefining the onco-drivers of tumor progression. *J. Exp. Clin. Cancer Res.* **39**, 215 (2020).

171. Mitsopoulos, C. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.* **49**, D1074–D1082 (2021).
172. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416.e11 (2018).
173. Therneau, T. M. *A Package for Survival Analysis in R*. <https://CRAN.R-project.org/package=survival> (2021).
174. Drawing Survival Curves using 'ggplot2' [R package survminer version 0.4.9]. (2021).
175. Bolker, B. & Robinson, D. Broom. mixed: Tidying methods for mixed models.
176. Wild, F. *lsa: Latent Semantic Analysis*. (2015).
177. Kuhn, M. caret: Classification and Regression Training. (2020).
178. Gray, B. cmprsk: Subdistribution Analysis of Competing Risks. (2020).
179. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A Lego System for Conditional Inference. *Am. Stat.* **60**, 257–263 (2006).
180. Pedersen, T. L. patchwork: The Composer of Plots. (2019).
181. van den Boogaart, K. G., Tolosana-Delgado, R. & Bren, M. compositions: Compositional Data Analysis. (2021).
182. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2020).
183. Coombes, K. R., Brock, G., Abrams, Z. B. & Abruzzo, L. V. Polychrome: Creating and Assessing Qualitative Palettes with Many Colors. *Journal of Statistical Software, Code Snippets* vol. 90 1–23 (2019).
184. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'. (2019).
185. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).
186. Wickham, H. Reshaping Data with the reshape Package. *Journal of Statistical Software* vol. 21 1–20 (2007).
187. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2021).
188. Attali, D. & Baker, C. ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements. (2019).
189. Wickham, H. & Seidel, D. scales: Scale Functions for Visualization. (2020).
190. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
191. Petukhov, V., van den Brand, T. & Biederstedt, E. ggrastr: Rasterize Layers for 'ggplot2'. (2021).
192. Lüdtke, D. sjPlot: Data Visualization for Statistics in Social Science. (2021).
193. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2020).
194. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. (2019).
195. Arnold, J. B. ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. (2019).
196. Henry, L., Wickham, H. & Chang, W. ggstance: Horizontal 'ggplot2' Components. (2019).
197. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2020).
198. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* vol. 4 1686 (2019).
199. Auguie, B. gridExtra: Miscellaneous Functions for 'Grid' Graphics. (2017).
200. Knaus, B. J. & Grünwald, N. J. VcfR: a package to manipulate and visualize VCF format data in R. *bioRxiv* 041277 (2016) doi:10.1101/041277.
201. Edwards, S. M. lemon: Freshing Up your 'ggplot2' Plots. (2020).
202. Garnier *et al.* viridis - Colorblind-Friendly Color Maps for R. (2021) doi:10.5281/zenodo.4679424.
203. Long, J. A. jtools: Analysis and Presentation of Social Scientific Data. <https://cran.r-project.org/package=jtools> (2019).

- End of document -