

Reporting Summary

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

Statistics

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

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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

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

Main repositories

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

Support repositories

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

github.com/markowetzlab/swgs-binsize

For the majority of analyses the free statistical software R (v3.6.1) was used. The following packages were used:

broom.mixed 0.3.3
 caret 6.0-83
 cmprsk 2.2-10
 coin 1.4-2
 compositions 2.0-1
 cowplot 1.0.0
 data.table 1.12.8
 DNACopy 1.64.0
 flexmix 2.3-15
 ggExtra 0.9
 ggplot2 3.1.1
 ggtrastr 0.2.3
 ggrepel 0.8.2
 ggthemes 4.2.0
 ggstance 0.3.3
 ggpubr 0.2.5
 gridExtra 2.3
 lemon 0.4.4
 jtools 2.0.3
 lsa 0.73.1
 MASS 7.3_51.5
 NMF 0.21.0
 patchwork 1.0.0
 PharmacGx 1.17.1
 Polychrome 1.3.1
 RColorBrewer 1.1-2
 reshape2 1.4.3
 rstatix 0.7.0
 scales 1.0.0
 sfsmisc 1.1-7
 sjPlot 2.8.9
 stringr 1.4.0
 survival 3.1-11
 survminer 0.4.8
 tidyverse 1.3.1
 vcfR 1.12.0
 viridisLite 0.3.0
 YAPSA 1.12.0

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

Data analysis

Signature derivation

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

Signature robustness

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

Simulations of cancer genomes, SBS signatures, single-feature vs multi-feature approaches

NMF (R package, v0.21.0)
 SigProfilerExtractor (Python, v1.1.4)

Association with biological covariates AND drug response and drug target analysis

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

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

Survival analysis

Univariate survival analysis (Kaplan-Meier estimates): survfit function from survival (R package, v.3.1-11)

Multivariate survival analysis (Cox proportional hazard models): coxph function from survival (R package, v.3.1-11)

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

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

Restricted data

- TCGA SNP6 Affymetrix arrays and copy number profiles

Paper: Beroukhi et al. (Nature, 2010)

URL to data:

TCGA Research Network: <https://www.cancer.gov/tcga>

Raw data: <https://portal.gdc.cancer.gov/legacy-archive/search/f>

Copy number segments: <https://portal.gdc.cancer.gov/repository>

Data access:

Part controlled (CEL files, raw data): Study Accession number (dbGaP): phs000178.v11.p8

Part open (copy number segments, processed data): ASCAT Github repository: <https://github.com/VanLoo-lab/ascat/tree/master/ReleasedData>

- PCAWG Copy number calls and Structural variants

Paper: Gerstung et al., Nature, 2020; Drento et al., Cell, 2021

URL to data:

US TCGA portion of PCAWG: <https://portal.gdc.cancer.gov/repository>

ICGC portion of PCAWG: <https://dcc.icgc.org/pcawg>

Data access:

Controlled (raw data):

Study Accession numbers:

For US TCGA portion of the PCAWG data: Phs000178.v11.p8 (dbGaP) -> https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8#restricted-access

For the ICGC portion of the PCAWG data: EGAS00001001692 (EGA) with an application to the ICGC-DACO -> <https://daco.icgc.org/>

Databases

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

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

- A snapshot of the Mitelman database can be accessed by the publicly available paper from Weaver and Cleveland: <https://www.sciencedirect.com/science/article/pii/S0955067406001554?via%3Dihub> A file summarising the relevant table 1 in the paper is found in our Github repository: https://github.com/markowetzlab/Dreus2022_CIN_Compendium/blob/main/Section%201%20-%204%20Signature%20Discovery/input/external/WeaverCleveland2006_Table1.txt

Unrestricted public data

TCGA

- Affymetrix SNP6.0 arrays: Beroukhi et al. (Nature, 2010)

- Gene expression: COSMIC database (Tate et al., NAR, 2019)

- Somatic exon SNVs in TCGA: Ellrott et al. (Cell Systems, 2018)

- Clinical data: TCGA Research Network (<https://www.cancer.gov/tcga>)

- Germline and somatic BRCA1/2 status for OV and BRCA patients: Wang et al. (Nature Genetics, 2017); Maxwell et al. (Nature Communication, 2017)

- Promoter hypermethylation of RAD51C and BRCA1 in TCGA-OV samples: TCGA Research Network (Nature, 2011); Sample IDs obtained from corresponding author.

- Telomere length, TelomereSignatureScore, Genomic status and expression of TERT/TERC, TERRA expression, Alternative Telomere Lengthening (ATRX/DAXX): Barthel et al. (Nature Genetics, 2017)

- Driver genes and annotations: Bailey et al. (Cell, 2018)

- Extrachromosomal DNA amplicons: Kim et al. (Nature Genetics, 2020)

- Tandem duplication identification and classification: Menghi et al., (Cancer Cell, 2018)

- HRD score, telomeric imbalance score, large-scale state transitions score, loss of heterozygosity score: Knijnenburg et al. (Cell Reports, 2018), Marquard et al. (Biomarker Research, 2015)

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

PCAWG

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

ICGC

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

Cell lines

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

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

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

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

Public but unpublished data

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size

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

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

Data exclusions

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

The initial set of arrays included 11,034 patients with 12,240 samples. First, we excluded noisy samples by filtering for the following three criteria:

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

Finally, the following three criteria were used for the exclusion of low quality arrays ($n = 189$):

- Tumour or normal MAPD < 0.75 : The median of absolute pairwise differences between adjacent probes (MAPD) is a per-microarray estimate of the variance, similar to the standard deviation or the interquartile range³⁶. Filtering for the MAPD removed noisy samples.
- Tumour and normal MAPD < 0.4 and fraction homozygous segments < 0.1 : Occasionally tumour samples are matched with the wrong

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

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

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

Replication

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

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

Randomization

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

Blinding

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

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging