nature research

Corresponding author(s): Constance Li
Paul C Boutros

Last updated by author(s): Jan 15, 2020

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 Authors & Referees and the Editorial Policy Checklist.

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Data and metadata were collected from International Cancer Genome Consortium (ICGC) members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at https://www.overture.bio. Other data collection software using in this effort, such as ICGC-specific portal user interfaces, are available upon request via contact@overture.bio.
Data analysis	The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v3.0, which allows for reuse and distribution. All statistical analyses and data visualization were performed in the R statistical environment (v3.4.3) using the BPG (v5.9.8)24 and car (v3.0-2) packages.

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/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

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium are described in the marker paper13 and available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially

identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. To access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorisation.
In addition, the analyses in this paper used a number of datasets that were derived from the raw sequencing data and variant calls (Supplementary Table 10). The individual data sets are available at Synapse (https://www.synapse.org/), and are denoted with synXXXXX accession numbers (listed under Synapse ID); all these datasets are also mirrored at https://dcc.icgc.org, with full links, filenames, accession numbers and descriptions detailed in Supplementary Table 10. Tumour histological classifications were reviewed and assigned by the PCAWG Pathology and Clinical Correlates Working Group (annotation version 9; syn10389158, syn10389164). Ancestry imputation was performed using an ADMIXTURE23-like algorithm by the PCAWG Germline Cancer Genome Working Group based on germline SNP profiles determined by whole-genome sequencing of the reference sample (syn4877977). The consensus somatic SNV and indel (syn7357330) file covers 2778 whitelisted samples from 2583 donors. Driver events were called by the PCAWG Drivers and Functional Interpretation Group (syn11639581). Consensus CNA calls from the PCAWG Structural Variation Working Group were downloaded in VCF format (syn8042988). Subclonal reconstruction was performed by the PCAWG Evolution and Heterogeneity Working Group (syn8532460). SigProfiler mutation signatures were determined by the PCAWG Mutation Signatures and Processes Working Group for single base substitution (syn11738669), doublet base substitution (syn11738667) and indel (syn11738668) signatures. Quality control measures data was provided by the PCAWG Technical Working Group (syn5864470).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	All data available from the PCAWG project was used with exclusions as described below. Cancer types with small sample sizes were ignored in tumour subtype-specific analyses
Data exclusions	Data from sex-specific cancers were excluded. Data also was excluded if annotation data describing sex was unavailable, as this information is critical to the study design.
Replication	In order to evaluate the performance of each of the mutation-calling pipelines and determine an integration strategy, the PCAWG Consortium performed a large-scale deep sequencing validation experiment. We selected a pilot set of 63 representative tumour/normal pairs, on which we ran the three core pipelines, together with a set of 10 additional somatic variant-calling pipelines contributed by members of the SNV Calling Working Group. Overall, the sensitivity and precision of the consensus somatic variant calls were 95% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement. Driver mutation, CNA and mutation density findings concur with previously reported findings. Sex-biases in non-coding mutation density, measures of tumour evolution and mutational signatures are intriguing and should be validated in future well-powered datasets with deep annotation to allow sufficient control for confounders.
Randomization	All relevant samples were included in the study.
Blinding	All participants were recruited into the study based only on clinical features. Blinding on mutational profiles and sex was not performed.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patient-by-patient clinical data are provided in the marker paper for the PCAWG consortium (Extended Data Table 1 of that manuscript). Demographically, the cohort included 1,469 males (55%) and 1,189 females (45%), with a mean age of 56 years (range, 1-90 years). Using population ancestry-differentiated single nucleotide polymorphisms (SNPs), the ancestry distribution was heavily weighted towards donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American and Australian projects. We consolidated histopathology descriptions of the tumour samples, using the ICD-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.
Recruitment	Patients were recruited by the participating centres following local protocols.
Ethics oversight	The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.