

Reporting Summary

Nature Portfolio 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 Portfolio 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 |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> 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> |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> 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) |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> 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> |
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> 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	Existing studies were used, so no software or code were used for data collection.
Data analysis	Statistical analysis were performed using R version 4.0.0 R bigsnpr package for LDpred2 Python based code for PRS-CSx https://github.com/getian107/PRScsx PRS generation plink2 score function Adjusted AUC were calculated using ROct R package Decision Curve Analysis were performed using R function stdca.R Survival analysis were performed using R packages survminer and survival Kaplan Meier plots were generated using R function ggsurvplot

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 Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The Summary-level data for the full set of Asian and European GWASs used in this study are available in the GWAS catalog under accession code GCST90129505 [<https://www.ebi.ac.uk/gwas/studies/GCST90129505>].

Genotype data of GERA participants who consented to having their data shared with dbGaP are available from dbGaP under accession phs000674.v2.p2. The complete GERA data are available upon successful application to the KP Research Bank. Genotype data of eMERGE participants are available from dbGaP under the accession number phs001616.v1.p1. For individual-level data, MEC, CCFR, The MD Anderson Colorectal Cancer Case Control Study, HCCS are deposited in dbGaP (phs000220.v2.p2, phs002733.v1.p1, phs002691.v1.p1, phs001193.v1.p1) and PLCO (phs001286.v3.p2). SCCS and CanCORS data can be accessed via websites <http://ors.southerncommunitystudy.org> and <http://outcomes.cancer.gov/cancers/>. For the remaining studies please contact the corresponding PIs: CR2&3 (Loic Le Marchand at loic@cc.hawaii.edu), Fukuoka, (Loic Le Marchand at loic@cc.hawaii.edu), Nagano, JPHC (Motoki Iwasaki at moiwasak@ncc.go.jp), UNC-Rectal (Temitope Keku at temitope_keku@med.unc.edu) and Basque Study (Prof Luis Bujanda at LUIS.BUJANDAFERNANDEZDEPIEROLA@osakidetza.eus). The 1000 Genomes phase 3 dataset (GRCh37) is available in PLINK2 binary format at PLINK 2.0 Resources (https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3). The PRS weight files generated by this study will be available in PGS catalog (<https://www.pgscatalog.org/>) upon publication of the manuscript.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Self-reported sex and gender from study questionnaire or electronic medical records at study entry

Population characteristics

GWAS summary statistics derived from European and Asian population origins, included 100,204 colorectal cancer cases and 154,587 controls. Validation data included participants from self-reported Asian, Black or African-American, hispanic and non-Hispanic White, racial and ethnic groups. The covariate-relevant population characteristics of the training and validation data set participants are given in Supplemental Table 1 and Table 1 respectively.

Recruitment

The study uses existing data. The recruitment of participants were done in previously conducted studies; see more details of these studies in the supplemental text and references.

Ethics oversight

All study protocols were approved by the relevant Institutional Review Boards, and informed consent was obtained from all study participants in accordance with the Helsinki accord.

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

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

We used all publicly available GWAS summary statistics for colorectal cancer cases and control for model building (100,204 cases and 154,587 controls) to maximize efficiency for deriving the best performing polygenic risk score. For validation data sets, we include all available individual-level data sets that have not been previously included in any of GWAS discoveries to ensure independent validation. Altogether, the validation data sets include 1,681-3,651 cases and 8,696-115,105 controls. Both the point estimates (e.g., AUC and odds ratio) and their standard errors are provided to quantify the performance of our proposed polygenic risk score (PRS) and its uncertainty. We demonstrated that the PRS was predictive of colorectal cancer risk with the 95% confidence intervals exclude the null (i.e. no prediction).

Data exclusions

Participants were excluded based on the previously established QC pipelines (e.g., genotyping failures). No exclusion was done in the analysis stage.

Replication

We assessed the performance of PRS using independent data sets which were not part of the training data sets and confirmed the PRS that we developed were predictive of colorectal cancer risk based on various evaluation metrics including AUC, odds ratio per PRS standard deviation, and net benefit analysis.

Randomization	The goal of this paper is to develop and validate polygenic risk score (PRS) for predicting colorectal cancer risk. There was not treatment or intervention to be evaluated. Therefore, randomization was not applicable.
Blinding	The study is observational and so no blinding is required.

Behavioural & social sciences study design

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

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

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

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

Location

Access & import/export

Disturbance

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

Clinical data

Dual use research of concern

Methods

n/a Involved in the study

ChIP-seq

Flow cytometry

MRI-based neuroimaging

Antibodies

Antibodies used

Validation

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines (See [ICLAC](#) register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Dating methods

Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

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

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

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

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Public health
<input checked="" type="checkbox"/>	<input type="checkbox"/> National security
<input checked="" type="checkbox"/>	<input type="checkbox"/> Crops and/or livestock
<input checked="" type="checkbox"/>	<input type="checkbox"/> Ecosystems
<input checked="" type="checkbox"/>	<input type="checkbox"/> Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
<input checked="" type="checkbox"/>	<input type="checkbox"/> Demonstrate how to render a vaccine ineffective
<input checked="" type="checkbox"/>	<input type="checkbox"/> Confer resistance to therapeutically useful antibiotics or antiviral agents
<input checked="" type="checkbox"/>	<input type="checkbox"/> Enhance the virulence of a pathogen or render a nonpathogen virulent
<input checked="" type="checkbox"/>	<input type="checkbox"/> Increase transmissibility of a pathogen
<input checked="" type="checkbox"/>	<input type="checkbox"/> Alter the host range of a pathogen
<input checked="" type="checkbox"/>	<input type="checkbox"/> Enable evasion of diagnostic/detection modalities
<input checked="" type="checkbox"/>	<input type="checkbox"/> Enable the weaponization of a biological agent or toxin
<input checked="" type="checkbox"/>	<input type="checkbox"/> Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*

Field strength *Specify in Tesla*

Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI Used Not used

Preprocessing

Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: Whole brain ROI-based Both

Statistic type for inference (See [Eklund et al. 2016](#)) *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a | Involved in the study
 Functional and/or effective connectivity
 Graph analysis
 Multivariate modeling or predictive analysis

Functional and/or effective connectivity *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*