Cell Genomics

CellPress
OPEN ACCESS

Perspective

The Mutographs biorepository: A unique genomic resource to study cancer around the world

Sandra Perdomo,[1,67] Behnoush Abedi-Ardekani,[1,67] Ana Carolina de Carvalho,[1,67] Aida Ferreiro-Iglesias,[1,67] Valérie Gaborieau,[1] Thomas Cattiaux,[1] Hélène Renard,[1] Priscilia Chopard,[1] Christine Carreira,[2] Andreea Spanu,[1] Arash Nikmanesh,[1] Ricardo Cortez Cardoso Penha,[1] Samuel O. Antwi,[3,4] Patricia Ashton-Prolla,[5,6] Cristina Canova,[7] Taned Chitapanarux,[8] Riley Cox,[9] Maria Paula Curado,[10] José Carlos de Oliveira,[11] Charles Dzamalala,[12]

(Author list continued on next page)

[1]Genomic Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France
[2]Evidence Synthesis and Classification Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France
[3]Division of Epidemiology, Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, FL, USA
[4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Jacksonville, FL, USA
[5]Experimental Research Center, Genomic Medicine Laboratory, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
[6]Post-Graduate Program in Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[7]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardio-Thoraco-Vascular Sciences and Public Health, University of Padua, Padova, Italy
[8]Department of Internal Medicine, Chiang Mai University, Chiang Mai, Thailand
[9]Ontario Tumour Bank, Ontario Institute for Cancer Research, Toronto, ON, Canada
[10]Department of Epidemiology, A.C. Camargo Cancer Center, São Paulo, Brazil
[11]Associação de Combate ao Câncer em Goiás Hospital Araújo Jorge Goiânia, Goiânia, Brazil
[12]University of Malawi College of Medicine, Blantyre, Malawi
[13]Regional Authority of Public Health, Banská Bystrica, Slovakia
[14]Departments of Surgery and Oncology, McGill University, Montreal, QC, Canada
[15]Early Cancer Institute, University of Cambridge, Cambridge, UK
[16]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic
[17]Mount Sinai Hospital; Ontario Institute for Cancer Research (OICR), Toronto, ON, Canada
[18]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD, USA
[19]Institute of Public Health & Preventive Medicine, 2nd Faculty of Medicine, Charles University, Prague, Czech Republic
[20]Department of Oncology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic
[21]National Cancer Institute, Bogotá, Colombia
[22]Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic
[23]Laboratory of Genetic Diagnostic, National Cancer Institute, Vilnius, Lithuania
[24]Department of Botany and Genetics, Institute of Biosciences, Vilnius University, Vilnius, Lithuania

(Affiliations continued on next page)

SUMMARY

Large-scale biorepositories and databases are essential to generate equitable, effective, and sustainable advances in cancer prevention, early detection, cancer therapy, cancer care, and surveillance. The Mutographs project has created a large genomic dataset and biorepository of over 7,800 cancer cases from 30 countries across five continents with extensive demographic, lifestyle, environmental, and clinical information. Whole-genome sequencing is being finalized for over 4,000 cases, with the primary goal of understanding the causes of cancer at eight anatomic sites. Genomic, exposure, and clinical data will be publicly available through the International Cancer Genome Consortium Accelerating Research in Genomic Oncology platform. The Mutographs sample and metadata biorepository constitutes a legacy resource for new projects and collaborations aiming to increase our current research efforts in cancer genomic epidemiology globally.

INTRODUCTION

Large-scale biomedical databases and resources help to promote advances in cancer research applications in diverse areas including cancer prevention, early detection, therapy, cancer care, and surveillance. The need to diversify the current knowledge on cancer genomics around the world requires the development and sustainability of large cancer biorepositories in different geographical regions with complete epidemiological data including demographics, lists

Elenora Fabianova,[13] Lorenzo Ferri,[14] Rebecca Fitzgerald,[15] Lenka Foretova,[16] Steven Gallinger,[17] Alisa M. Goldstein,[18] Ivana Holcatova,[19,20] Antonio Huertas,[21] Vladimir Janout,[22] Sonata Jarmalaite,[23,24] Radka Kaneva,[25] Luiz Paulo Kowalski,[10,26] Tomislav Kulis,[27,28] Pagona Lagiou,[29] Jolanta Lissowska,[30] Reza Malekzadeh,[31] Dana Mates,[32] Valerie McCormack,[33] Diana Menya,[34] Sharayu Mhatre,[35] Blandina Theophil Mmbaga,[36] André de Moricz,[37] Péter Nyirády,[38] Miodrag Ognjanovic,[39] Kyriaki Papadopoulou,[40] Jerry Polesel,[41] Mark P. Purdue,[42] Stefan Rascu,[43] Lidia Maria Rebolho Batista,[44] Rui Manuel Reis,[44,45] Luis Felipe Ribeiro Pinto,[46] Paula A. Rodríguez-Urrego,[47] Surasak Sangkhathat,[48] Suleeporn Sangrajrang,[49] Tatsuhiro Shibata,[50,51] Eduard Stakhovsky,[52] Beata Świątkowska,[53] Carlos Vaccaro,[54] Jose Roberto Vasconcelos de Podesta,[55] Naveen S. Vasudev,[56] Marta Vilensky,[57] Jonathan Yeung,[58] David Zaridze,[59] Kazem Zendehdel,[60] Ghislaine Scelo,[61] Estelle Chanudet,[62] Jingwei Wang,[63] Stephen Fitzgerald,[63] Calli Latimer,[63] Sarah Moody,[63] Laura Humphreys,[63] Ludmil B. Alexandrov,[64,65,66] Michael R. Stratton,[63] and Paul Brennan[1,*]

[25]Molecular Medicine Center, Department of Medical Chemistry and Biochemistry, Medical Faculty, Medical University of Sofia, Sofia, Bulgaria
[26]University of São Paulo Medical School, São Paulo, Brazil
[27]Department of Urology, University Hospital Center Zagreb, Zagreb, Croatia
[28]University of Zagreb School of Medicine, Zagreb, Croatia
[29]National and Kapodistrian University of Athens, Athens, Greece
[30]The Maria Sklodowska-Cure National Research Institute of Oncology, Warsaw, Poland
[31]Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran
[32]Occupational Health and Toxicology, National Center for Environmental Risk Monitoring, National Institute of Public Health, Bucharest, Romania
[33]Environment and Lifestyle Epidemiology Branch, International Agency for Research on Cancer (IARC/WHO), Lyon, France
[34]Moi University, School of Public Health, Eldoret, Kenya
[35]Division of Molecular Epidemiology and Population Genomics, Centre for Cancer Epidemiology, Tata Memorial Centre, Mumbai, India
[36]Department of Surgery, Santa Casa School of Medicine, São Paulo, Brazil
[37]Kilimanjaro Clinical Research Institute, Kilimanjaro Christian Medical Centre & Kilimanjaro Christian Medical University College, Moshi, Tanzania
[38]Semmelweis University, Budapest, Hungary
[39]IOCPR- International Organization for Cancer Prevention and Research, Serbia, Belgrade
[40]Hellenic Cooperative Oncology Group (HeCOG), Athens, Greece
[41]Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Aviano, Italy
[42]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
[43]Urology Department, "Carol Davila" University of Medicine and Pharmacy – "Prof. Dr. Th. Burghele" Clinical Hospital, Bucharest, Romania
[44]Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
[45]Life and Health Sciences Research Institute (ICVS), School of Medicine, Minho University, Braga, Portugal
[46]Brazilian National Cancer Institute, Rio de Janeiro, Brazil
[47]Pathology Department, Hospital Universitario Fundación Santa Fe de Bogotá, Bogotá, Colombia
[48]Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Hat Yai, Thailand
[49]National Cancer Institute, Bangkok, Thailand
[50]Laboratory of Molecular Medicine, The Institute of Medical Science, The University of Tokyo, Minato-ku, Japan
[51]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Japan
[52]National Cancer Institute, Kiev, Ukraine
[53]Department of Environmental Epidemiology, Nofer Institute of Occupational Medicine, Łódź, Poland
[54]Instituto Medicina Traslacional e Ingenieria Biomedica - CONICET, Buenos Aires, Argentina
[55]Hospital Santa Rita AFECC - Associação Feminina de Educação e Combate ao Câncer, Vitoria, Brazil
[56]Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
[57]Instituto de Oncología Angel Roffo, Universidad de Buenos Aires, Buenos Aires, Argentina
[58]University Health Network, Toronto, ON, Canada
[59]Clinical Epidemiology, N.N. Blokhin National Medical Research Centre of Oncology, Moscow, Russia
[60]Cancer Research Center, Cancer Institute, Tehran University of Medical Sciences, Tehran, Iran
[61]Observational & Pragmatic Research Institute Pte., Ltd., Singapore, Singapore

(Affiliations continued on next page)

of relevant environmental exposures, and detailed clinical information.

Whole-genome sequencing (WGS) of tumor-normal matched pairs is a powerful method to determine the diversity and complexity of somatic and germline mutations for both understanding the etiology and revealing diagnosis and treatment opportunities in patients with cancer. Despite the large amounts of publicly available WGS data, generated from patients with cancer as part of the global PanCancer Analysis of Whole Genomes (PCAWG) project[1] (n = 3,109) of the International Cancer Genome Consortium (ICGC), the Hartwig Foundation (n = 5,520),[2] and, more recently, from the Genomics England's 100,000 Genomes Project (n = 12,222),[3] these efforts have focused almost exclusively on patients from

CellPress
OPEN ACCESS

[62]Department of Pathology, Radboud University Medical Centre, Nijmegen, the Netherlands
[63]Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Cambridge, UK
[64]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA
[65]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[66]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA
[67]These authors contributed equally
*Correspondence: brennanp@iarc.who.int
https://doi.org/10.1016/j.xgen.2024.100500

Europe, North America, and Australia, with a limited representation of cancers from Asian and African countries (Figure 1). Inclusion of more diverse populations of patients from other geographical regions in cancer genomic studies still lags behind,[4] and emerging findings from comparative studies of population diversity in cancer genomics[5,6] have established the necessity to expand diverse genetic and epidemiological data.

The Mutographs project is a Cancer Grand Challenges partnership funded by Cancer Research UK (CRUK: C98/A24032) with the primary objective of elucidating the causes of major global geographical and temporal differences in cancer incidence through mutational signature analysis. The project is generating mutational signatures and additional genomic descriptive analyses from WGS of thousands of paired (tumor/blood) samples from patients with cancer diagnosed with selected cancer types in 30 countries across 5 continents (Figure 2). The Mutographs study also consolidates an international research network working on cancer genomic epidemiology and exemplifies how genomic studies in cancer can promote scientific inclusion and equity through international collaboration.

GENOMIC EPIDEMIOLOGY APPROACHES FOR IDENTIFYING NEW CAUSES OF CANCER GLOBALLY

Differences in cancer incidence between populations cannot be uniquely attributable to endogenous mutagenic processes. For instance, evidence from migrant studies[7] and recent time trends[8–11] show that genetic susceptibility cannot sufficiently explain these differences, indicating that lifestyle and environmental factors should also be responsible.

Traditional epidemiological studies mostly based on large retrospective case-control analysis have exhausted the possibilities of finding or confirming new potential causes of cancer because data are outdated or non-existent for current and/or relevant exposures.[12,13] The more recently available large prospective, population-based cohorts and cohort consortia have improved the resolution of cancer etiological findings by improving study quality for rare exposures, linking data across sources such as electronic health records, tumor biobanks, cancer registries, geospatial data, and mobile data, among others, and independently validating previous findings.[14] However, these consortia are far from being representative of many geographical areas and will fail to identify unknown exposures possibly relevant for certain regions. The inclusion of state-of-the-art genomic studies complementary to well-defined epidemiological study designs and extensive data collection redefines the new era of genomic epidemiology studies in cancer research and can help to identify unknown causes of cancer worldwide.[15,16]

MUTOGRAPHS RATIONALE: ELUCIDATING GLOBAL DIFFERENCES IN CANCER INCIDENCE USING GENOMIC EPIDEMIOLOGY

The Mutographs project focused on investigating the causes of large international differences in incidence and mortality for several cancers that are still poorly understood. Five initial cancer types, esophageal squamous cell carcinoma (ESCC), renal cell carcinoma (RCC), colorectal cancer (CRC), pancreatic ductal adenocarcinoma (PDAC), and gastroesophageal junction adenocarcinoma (GEJ), including both esophageal adenocarcinoma (EAD) and adenocarcinomas of gastric cardia, were selected based on the following criteria: (1) cancers with the highest differences in incidence across geographical regions, (2) cancers accounting for more than 10% of new cases and about 20% of deaths, and (3) cancers for which prevalence of known risk factors (i.e., smoking, alcohol, and obesity) do not fully explain the large geographical and temporal differences across regions (Figure 3).

For instance, CRC, RCC, and PDAC cancer cases show a similar distribution across geographical regions, being most common in Central Europe (particularly Czech Republic), North America, and East Asia (especially Japan and South Korea) and relatively rare in Africa and certain parts of Asia.[18] The known common risk factors for these three cancers are obesity[19] and tobacco smoking, although their effects are modest (~50% increased risk). Other suspected risk factors include dietary components such as animal protein, processed meat,[20] and alcohol consumption for CRC[21]; exposure to trichloroethylene,[22] aristolochic acid,[23,24] and per- and polyfluoroalkyl substances for RCC[25,26]; and clinical conditions such a hypertension and diabetes for RCC[27,28] and PDAC,[29] respectively.

Esophageal cancer, in particular ESCC, is an example of a cancer with differences in incidence within regions or countries. High ESCC rates are found in localized populations, including northeastern Iran, north and central China,[30] parts of Africa, and southern Brazil.[31] Established risk factors include tobacco, opium,[32] and alcohol,[21] but population attributable fractions vary between regions. Additionally, there is strongly suggestive evidence for a role of the consumption of very hot beverages[32] (tea/coffee/porridge in Africa, tea in Iran, maté in the south of Brazil) and a nutritionally deficient diet and exposure to polycyclic aromatic hydrocarbons from diverse sources, such as indoor biomass combustion.[33]

There has been a rapid increase of EAD in recent decades more evident among men and in specific populations,[34]

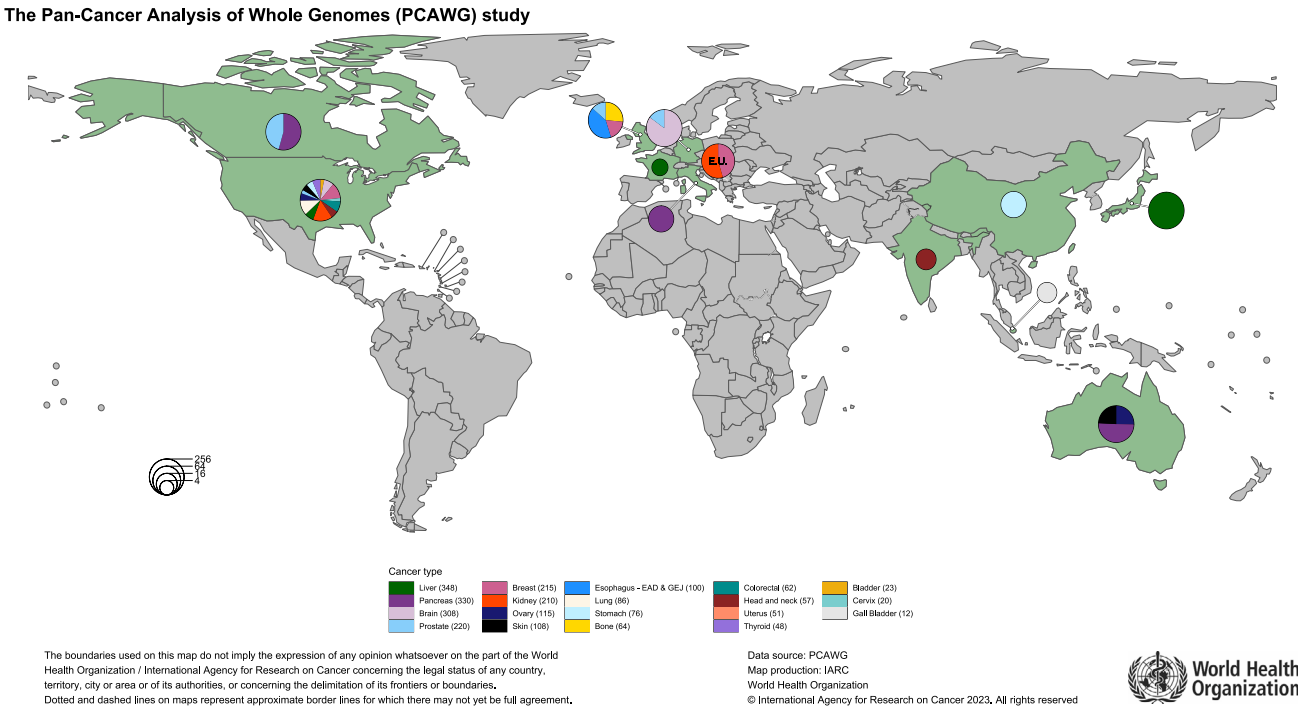


Figure 1. Geographical distribution of Pan-Cancer Analysis of Whole Genomes (PCAWG) study

particularly in Western Europe, North America, Australia, and in the Golestan province of Iran. The reasons for this particular geographical distribution are still unknown.[35]

Three additional cancer types were subsequently integrated in the study based on marked regional incidence differences and/or specific exposures of interest: (1) head and neck cancer (HNC), including cases identified in high- and intermediate-incidence countries in Europe[36] and South America[37] and oversampled for cases without reported tobacco and/or alcohol consumption; (2) urinary bladder cancer (UBC), with a selection of cases diagnosed in the Kerman province in Iran with and without documented consumption of opium,[38] recently classified as a carcinogenic substance (group I)[39]; and (3) gallbladder cancer (GBC), a cancer with the highest incidence rates in countries in South America and India and for which causes are poorly understood. Cases were selected from different regions in India where GBC is more common in women in the north, northeastern, and east (e.g., in the Kamrup urban district, incidence of GBC is 6.4 per 100,000 for men and 12.1 per 100,000 for women) compared to the southern part of India (in Mumbai, incidence is 1.8 per 100,000 for men and 2.8 per 100,000 in women)[40] (Figures 2 and 3).

The eight cancer sites included in the Mutographs study are also covered in the PCAWG study. However, Mutographs provides a larger number of cases for these cancer sites (4,397 cases already sequenced in Mutographs vs. 794 in PCAWG) and a broader geographic representation with 17 countries not explored in PCAWG from additional regions in eastern Europe, South America, South and East Asia, and Eastern Africa (Figure 2).

BUILDING UP A GLOBAL BIOREPOSITORY: CONSIDERING THE LOCAL PERSPECTIVE AND EXPERIENCE IN GLOBAL MULTICENTER STUDIES

The International Agency for Research on Cancer (IARC/WHO) promotes global collaboration in cancer research through the coordination of research across countries and organizations[41] and convening multidisciplinary expertise. With this long-standing track record and extensive international studies and network of partners, the Mutographs study was in the position to bring together existing studies as well as initiate *de novo* studies. Collaborating centers were selected among academic institutions, university hospitals, national cancer institutions, and private and public hospitals with experience in patient recruitment and sample and data acquisition for representative samples of patients with cancer. Through this international network of 50 institutional collaborators in 30 countries, between 2018 and 2022, the Mutographs team harmonized biospecimen and epidemiologic data on 7,808 cancer cases. Of the 7,808 cases, 3,023 (39%) were newly recruited, and 4,785 (61%) were based upon existing metadata and selected biorepositories.

A UNIFIED PROTOCOL FOR CASE SELECTION AND PROSPECTIVE RECRUITMENT WITH EXTENDED EXPOSURE DATA

Sample/data collection and inclusion criteria

Participant centers were included in the Mutographs project under two different scenarios or a combination of both.

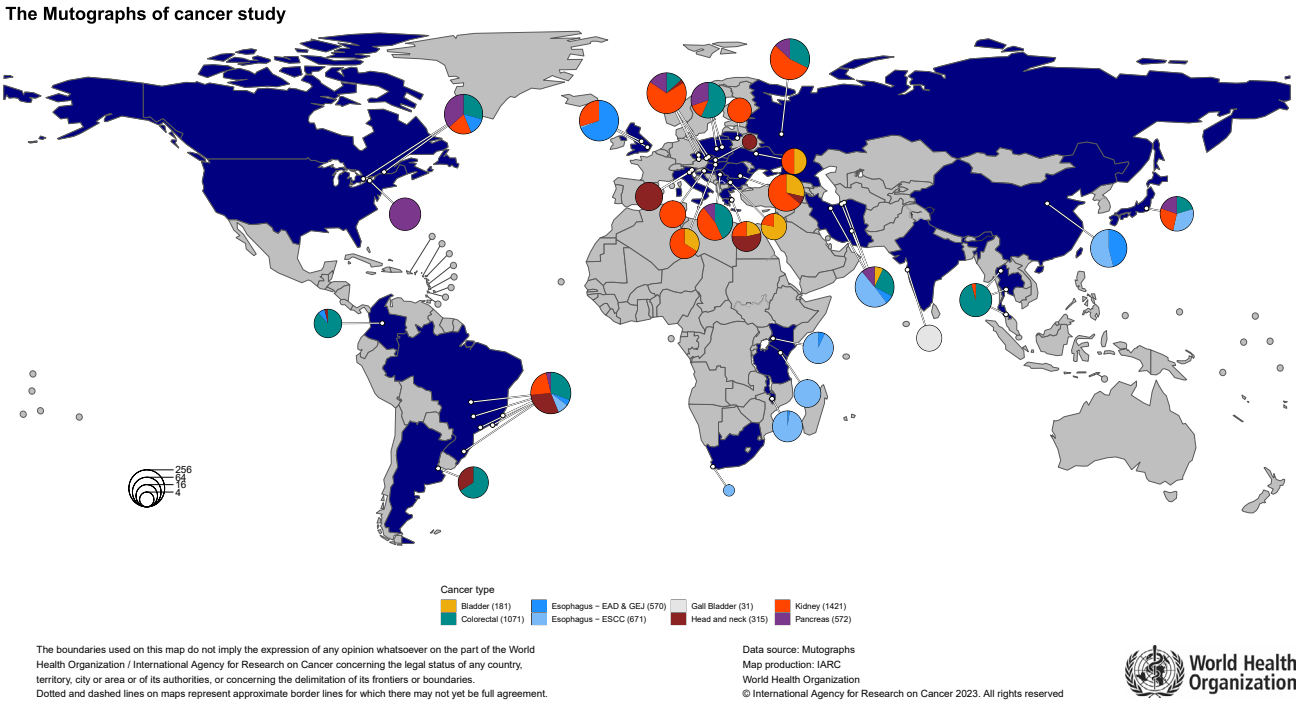


Figure 2. Geographical distribution of Mutographs cancer studies with complete whole-genome sequencing

Mutographs cases correspond to those that passed the quality controls for eligibility, tissue and blood processing, and sequencing. White circles represent cities where cases were recruited: Argentina: Buenos Aires; Brazil: Barretos, Goiania, Porto Alegre, Rio de Janeiro, Sao Paulo, and Vitoria; Bulgaria: Sofia; Canada: Montreal and Toronto; China: Shanxi; Colombia: Bogotá; Croatia: Zagreb; Czech Republic: Brno, Ceske Budejovice, Olomouc, and Prague; Greece: Athens; Hungary: Budapest; India: Mumbai; Iran: Gonbad, Gorgan, Kerman, and Tehran; Italy: Aviano and Padova; Japan: Tokyo; Kenya: Eldoret; Lithuania: Vilnius; Malawi: Blantyre; Poland: Lodz and Warsaw; Romania: Bucharest; Russia: Moscow; Serbia: Belgrade; Slovakia: Banska Bystrica; South Africa: Cape Town; Tanzania: Moshi; Thailand: Bangkok, Chiang Mai, and Hat Yai; Ukraine: Kiev; UK: Cambridge and Leeds; and US: Rochester.

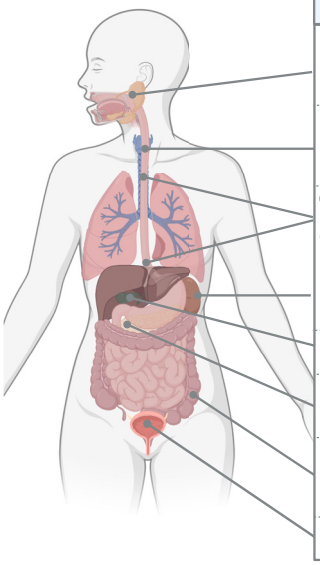
Scenario 1: ongoing or retrospective studies and biorepository collections not being previously sequenced/analyzed, published, or included in other international genomic initiatives such as PCAWG.[1] 31 centers provided patients that fulfilled the required criteria for inclusion (as described below).

Scenario 2: prospective collection of newly diagnosed patients. 19 centers prospectively recruited a sample of representative patients per cancer site. Patients were excluded if they had any condition that could interfere with their ability to provide informed consent or if there were no means of obtaining adequate tissues as per the protocol requirements. Ethical approvals were first obtained from each local research ethics committee and federal ethics committee when applicable, as well as from the IARC ethics committee.

Dedicated standard operating procedures (SOPs) were designed by IARC/WHO following guidelines from the ICGC to harmonize exposure, lifestyle, pathological, and clinical information from all cases to be included in the Mutographs project. The inclusion criteria for patients were that they were at least 18 years of age; had a confirmed diagnosis of primary tumors from the list of cancer sites eligible for the study; had no prior treatment; had an availability of fresh frozen tumor and, if possible, non-tumor fresh frozen tissue (FFT) and blood samples; had an availability of core epidemiological and clinical data (retrospectively or prospectively collected); and had ethics approval and consent for genetic studies and data sharing. For all patients prospectively included, after informed consent was obtained, anthropometric measures were taken, together with relevant information regarding medical and familial history. Blood samples were drawn, and a 30 min questionnaire was administered by a trained interviewer to collect complementary lifestyle and environmental information following the Mutographs SOPs.

Large sample collection

Fresh frozen tumor tissue and blood were collected from all cases as the most suitable samples for genomic studies. Non-tumor adjacent tissue was available for 53% of the cases. Collection of oral rinses and urine was also included for a subset of the HNC and UBC cohorts, respectively. Blood collection in EDTA tubes consisted in 10 mL preserved as whole blood. When feasible, blood samples were immediately processed into buffy coat, plasma, and red blood cells, followed by storage at −80°C. These collections resulted in more than 85,000 biological samples currently stored at the IARC/WHO BioBank (https://ibb.iarc.fr). Tumor and non-tumor tissues were collected before any treatment and while avoiding routine care disruption. Unless surgery was the first line of treatment, mucosal biopsies were collected. Tumor and non-tumor tissues were snap frozen in liquid nitrogen or promptly preserved in RNAlater (RNAprotect Tissue Tubes, QIAGEN). Laboratory information management



Organ sites	Known or suspected risk (R) factors	High incidence regions (male/female ASR)	Low incidence regions (male/female ASR)
Head and neck	**Alcohol, smoking, HPV**	Romania (33.4/2.9) Slovakia (30.8/3.2) Brazil (15.9/3.2) Czech Republic (15.6/4.9)	Iran, Tehran (0.9/0.9) Russia (0.8/0.1) Poland (0.5/0.1)
Oesophagus (ESCC)	**Smoking, alcohol, radiation**, hot drinks, opium	Kenya, Nairobi (14.7/15.2) Iran, Golestan (23.2/ 18.8) China, Huaiyin District (58.8/29.0)	United Kingdom (9.8/3.9) Japan (10.5/1.6) Brazil (8.1/2.1)
Oesophagus(EAD) and Gastroesophageal junction	GORD, obesity	United Kingdom (7.2/1.4) Iran, Golestan (29/6) Canada, Ontario (3.0/0.4) Brazil (2.0/0.6)	Iran, Tehran (0.9/0.9) Russia (0.8/0.1) Poland (0.5/0.1)
Kidney	**Obesity, smoking, trichloroethylene, AA, hypertension**	Czech Republic (22.1/9.9) Slovakia (15.1/7.5) Lithuania (20.5/8.2)	Japan (7.8/3.0) Brazil, Sao Paulo (6.5/3.1)
Gallbladder	**Gallstones, smoking**	Chile (7.2/1.4) India, Kamrup district (6.4/12.1)	Argentina (0.9/1.7) India, Mumbai (1.8/2.8)
Pancreas	**Smoking, obesity**, diabetes	Czech Republic (11.6/ 7.5) Slovakia (11.2/6.4) Japan, Miyagi (10.5/6.2)	Iran, Golestan (2.8, 1.0) Brazil, Sao Paulo (5.2/4.2)
Colorectum	Red meat, **processed meat, alcohol, obesity**	Canada, Ontario (40.7/27.5) Czech Republic (57.0/28.5) Japan (42.1/23.5) Brazil (29.5/26.6)	Iran, Golestan (15.8/12.4) Thailand, Chiang Mai (15.8/11.8) Colombia (13.4/12.5)
Bladder	**Smoking, opium, arsenic, AA**	Croatia (20.4/5.1)	Iran, Golestan (9.9/2.9)

Figure 3. Summary of cancer subsites included in the Mutographs biorepository by known and suspected risk factors and differences of incidence by regions

Age-standardized rates (ASRs) retrieved from *Cancer Incidence in Five Continents*, *Vol. XI*.[17] AA, aristolochic acid; HPV, human papillomavirus. Created with BioRender.com.

systems at local recruiting centers were used to keep track of samples collected and processed before being transferred to the IARC/WHO BioBank.

Detailed environmental exposure and clinical questionnaire: Diversifying local and regional exposure data

The focus on evaluation of exposures associated with recognized and plausible risk factors makes this biorepository a valuable source of detailed and comparable epidemiological metadata from patients with cancer diagnosed with the same tumor site but from diverse geographical regions. We compared the available exposure information among patients from cancer sites included in both PCAWG and Mutographs to estimate the extent of the project metadata. PCAWG includes information on tobacco smoking and alcohol consumption for three cancer sites. No additional exposure information was publicly accessible (Table 1). Data for those two exposures are partially available. History of tobacco smoking consumption is available for 30% of patients with PDAC from Canada, 20% of patients diagnosed with oral cavity (HNC) in India, and 80% of patients with EAD from the UK. History of alcohol consumption is absent in more than 70% of these cancer sites in PCAWG. In contrast, tobacco and alcohol consumption is documented in 97% and 81% of the Mutographs patients for the three cancer sites (Table 1). Mutographs has retrieved information on environmental exposures and risk factors for up to 90% of patients from all eight cancer sites using the following methodology.

Data from prospectively recruited patients were collected using a centralized database developed in the REDCap platform,[42] and IARC/WHO harmonized all retrospective data received as previously described.[43,44] The following core epidemiological and clinical data were required from all participants in the study: (1) demographic details (age, sex, ethnic origin, city and place of residence, and educational status); (2) history of tobacco use, including frequency and intensity; (3) history of alcohol consumption, including frequency and intensity; (4) anthropometric data (height, weight at diagnosis); (5) medical history of diabetes, hypertension, and acid reflux/heartburn for PDAC, RCC, and GEJ, respectively; and (6) consumption of hot drinks (for ESCC) or red and processed meat (for CRC). In addition to the data collected under the core variables mentioned above, all prospectively and partially retrospectively recruited patients provided information on oral health, physical activity, occupational exposures, and family history of cancer. We also collected information on the following regional and/or population-specific exposures.

(1) Opium consumption including route of administration, frequency, and intensity for patients with cancer recruited in Iran.
(2) Consumption of traditional South American maté including frequency, temperature, and intensity for patients from and/or recruited in the south of Brazil and in Argentina.
(3) Residential history and consumption of herbal remedies as possible sources of aristolochic acid exposure for patients with RCC from Romania, Serbia, Bulgaria, Croatia, Hungary, Greece, and Ukraine.

Clinical follow-up information up to 3 years after cancer diagnosis was retrieved, if possible, from clinical charts from retrospective collections, and additional information for up to 5 years is being collected from the prospectively recruited patients. IARC/WHO

Table 1. Distribution of PCAWG and Mutographs cancer cases with smoking and alcohol consumption information

	Esophagus		Head and neck (HNC)		Pancreas (PDAC)	
Characteristic	Mutographs (GEJ)	PCAWG (EAD)	Mutographs	PCAWG	Mutographs	PCAWG
Total	570	100	315	57	572	330
Countries	Brazil, China (Shanxi), Iran, Japan, Kenya, Malawi, Tanzania, UK	UK	Argentina, Brazil, Colombia, Czech Republic, Greece, Italy, Romania, Slovakia	India, US	Brazil, Canada, Czech Republic, Iran, Poland, Russia, Serbia, UK, US	Australia, Canada
Sex assigned at birth						
Female (%)	104 (18)	14 (14)	72 (23)	10 (18)	291 (51)	152 (46)
Male (%)	466 (82)	86 (86)	243 (77)	47 (82)	281 (49)	176 (53)
Unknown (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.6)
Age at diagnosis (years), median (IQR)	67 (60.0, 73.0)	70 (62.5, 76.0)	59 (50.0, 68.0)	53 (42.0, 62.0)	66 (59.1, 72.0)	65 (56.0, 73.0)
Unknown (%)	13.0 (2.3)	1.0 (1.0)	0.0 (0.0)	0.0 (0.0)	6.0 (1.0)	1.0 (0.3)
Tobacco status (%)						
Current smoker	159 (28)	20 (20)	152 (48)	8 (14)	122 (21)	5 (1.5)
Ever smoker	2 (0.4)	0 (0)	0 (0)	0 (0)	24 (4.2)	0 (0)
Ex-smoker	190 (33)	48 (48)	93 (30)	0 (0)	145 (25)	29 (8.8)
Never	171 (30)	20 (20)	70 (22)	5 (8.8)	280 (49)	61 (18)
Unknown	48 (8.4)	12 (12)	0 (0)	44 (77)	1 (0.2)	235 (71)
Alcohol status (%)						
Current drinker	107 (19)	0 (0)	122 (39)	0 (0)	98 (17)	0 (0)
Ever drinker	46 (8.1)	27 (27)	21 (6.7)	6 (11)	166 (29)	30 (9.1)
Ex-drinker	20 (3.5)	0 (0)	99 (31)	0 (0)	28 (4.9)	0 (0)
Never	126 (22)	0 (0)	73 (23)	7 (12)	251 (44)	11 (3.3)
Unknown	271 (48)	73 (73)	0 (0)	44 (77)	29 (5.1)	289 (88)

EAD, esophageal adenocarcinoma; GEJ, gastroesophageal junction adenocarcinoma; HNC, head and neck cancer; PDAC, pancreatic ductal adenocarcinoma; PCWAG, the global PanCancer Analysis of Whole Genomes.

harmonized all retrospective data. All data were de-identified locally through the use of a dedicated alpha-numerical identifier system before being transferred to IARC/WHO central database.

Centralized expert pathology review

Diagnostic pathology departments from participating centers provided diagnostic details on morphology and histology of patients through standard abstract forms, together with a representative hematoxylin and eosin (H&E)-stained slide of formalin-fixed, paraffin-embedded (FFPE) tumor tissues whenever possible. For all patients, to reconfirm the original histology, IARC/WHO centralized the entire pathology workflow on FFT tumors and coordinated their digital pathology examination included in the study, as well as FFPE sections when available, via a web-based report completed by a dedicated expert panel for each cancer site. High-resolution images of FFT tumors were randomly assigned to panel members, all blind to the original diagnosis. In addition to diagnosis and confirmation of tumor type, the percentage of viable cellular elements (tumor, inflammatory, and other non-tumor cells) and necrosis were recorded. 17% of randomly selected H&E slides underwent two independent pathology evaluations. A minimum of 50% viable tumor cells were required for eligibility to WGS. The percentage of processed cases with less than 50% tumor content was 18% for UBC, 20% for RCC, 31% for CRC, 36% for GBC, 43% for HNC, 45% for ESCC, and 49% for GEJ. Tumor enrichment procedures were applied, when possible, by laser capture microdissection (LCM) of the unwanted non-tumoral area. Approximately half of the GEJ cases and 96% of PDAC cases underwent LCM to enrich tumor cellularity.

Extraction of DNA and quantification from tumor and paired blood was centrally conducted at IARC/WHO[43] and is stored for subsequent analyses.

Out of the 7,808 recruited cancer cases, 4,400 were successfully processed at IARC/WHO, passed the pathological quality control metrics, and were sent to the Wellcome Sanger Institute for paired WGS and primary mutational signature analyses. 655 cases are yet to be processed and will be evaluated through our pathology pipeline and stored in the study biorepository for future studies.

SEQUENCING ANALYSES AND CURRENT PUBLICATIONS

The data sequencing pipeline has been developed and validated by the Wellcome Sanger Institute as previously described.[43]

WGS (150 bp paired end) is performed on the Illumina NovaSeq 6000 platform with target coverage of 40× for tumors and 20× for matched non-tumor tissues. Cases are excluded if coverage is below 30× for tumors or 15× for non-tumor tissue.

The data analysis workflow focuses on 4 areas: (1) the characterization of the tumor genome for each sample with specific data generated on driver genes, copy-number profiles, evaluation of tumor mutation burden, structural rearrangements, and other cancer-specific information such as the presence of viral and/or bacterial sequences for specific cancer sites; (2) the extraction and attribution of mutational signatures based on base substitutions (single and double), insertions/deletions, copy-number variants, and chromosomal rearrangements; (3) analyses highlighting possible contributions of germline variants and ancestry distribution[45] to the mutational signature profiles and associated exposures; and lastly, (4) associations between somatic genomic profiles and epidemiological data focused mainly on recognized and plausible risk factors. In 2021, the analysis on 552 patients with ESCC from eight countries with varying incidence rates was completed and showed a high prevalence of APOBEC signatures in all cases, as well as specific mutation signatures linked to opium and alcohol consumption, and homologous DNA repair deficiency.[43] Analysis of 962 patients with clear cell RCC is ongoing, and preliminary results are shedding light on the contribution of environmental causes on the high risk of this cancer in Central Europe and Japan.[44] Most of the sequencing and analysis efforts are now focused on cases of HNC, CRC, PDAC, and GEJ. By early 2023, we completed the sequencing of 2,777 matched-normal cancer genomes, and these samples are undergoing bioinformatics cancer genomics analysis. Data from these cases have been released to the Mutographs teams for subsequent combined analysis.

DATA REPOSITORY AND SHARING

A general description of the Mutographs project is available on the project website: https://www.mutographs.org. WGS data and patient metadata after analyses are being deposited and made publicly available via the European Genome Phenome Archive (EGA), currently associated with studies EGAS00001003542 and EGAS00001002725. All algorithms and codes used for genomic and epidemiological analysis and figures are publicly available with repositories noted in the respective publications.[43,44]

In addition, Mutographs is one of the participating programs of the ICGC Accelerating Research in Genomic Oncology (ARGO).[46] Therefore, all genomic, exposure, and clinical data will be publicly available through the ICGC ARGO data platform after agreement of the participating centers. An online catalog of the Mutographs biorepository is under development and will allow the broader research community to propose additional projects and/or analyses beyond the scope of Mutographs.

ONGOING AND FUTURE INITIATIVES FOR THE MUTOGRAPHS BIOREPOSITORY

The study of the mutational signatures operative in the genomes of patients with cancer around the world will generate a comprehensive catalog of the mutational processes that cause human cancer. An increasing number of signatures of different mutation classes are being reported, and correlations are being drawn to various exposures and/or endogenous factors. Experimental validation of signatures linked to specific exposures in human sample collections adds valuable information to establish causality.[47] Therefore, it is fundamental to gain a mechanistic understanding of how mutational signatures arise through experimental exploration.

To completely understand the etiology of cancer and to apply this knowledge to cancer prevention,[15] analysis of non-cancer tissues from patients with cancer and patients with benign or preneoplastic conditions[48] can provide insight into background mutational processes in healthy cells[49] and into the effects of suspected mutagens and exposures prior to the development of symptomatic lesions and, eventually, the diagnosis of cancer.[50–52] PROMINENT (CRUK: CGCATF-2021/100007, NIH: 1OT2CA278681-01) is a recently awarded Cancer Grand Challenges project aiming to detect and characterize mutagenic and promoting exposures before cancer develops using human tissue, mouse, and organoid models. Combined multiomics approaches are being used to understand the distributions of mutations and early neoplastic clones in non-tumor tissues from patients with cancer included in the Mutographs biorepository. Further investigation of mutagenic processes in non-tumor tissues using non-invasive sources of tissue (i.e., blood, urine, and nasopharyngeal, buccal, cervical, and anal swabs) will allow us to easily identify and monitor carcinogenic exposures and ultimately determine how these influence clinical and epidemiological patterns of cancer development. The resources created by Mutographs should also allow for subsequent studies addressing poorly understood cancer-related questions and shed light on the current and future challenges in cancer research.

BEYOND MUTOGRAPHS: EXPANDING GENOMIC EPIDEMIOLOGICAL REPOSITORIES AND INCLUSIVE RESEARCH COLLABORATIONS

The Mutographs study is an example of novel genomic initiatives needed to expand our understanding of causes and processes related to cancer onset on a global scale. Some of the key aspects that contributed to the successful creation of such a large-scale cancer biorepository and that we suggest should be applied to similar initiatives in the future include the following.

Establishing long-lasting research collaborations and standardized protocols: a fundamental pillar in any international collaborative project involves complying with a detailed and unique recruitment protocol for patient selection and sample collection while adapting to the specific local or regional context for an effective implementation of the study protocol. Such a balance was achieved in Mutographs by close communication and follow up with the different institutions throughout the duration of the project. The Mutographs study team investigated the local research needs in terms of available personnel, minimum infrastructure requirements, and institutional procedures for patient approach, routes of diagnosis, and treatment. Questionnaires were revised to include relevant exposures, adapt the questions' wording, and avoid possible sensitive questions. Center-specific adjustments were included in the protocols if necessary.

Research agreements as well as material and data transfer agreements were better established after close discussions with collaborators to comply with legal and data protection requirements in each country, the research institutions, and the funders. Essentially, publication policies and ownership of data and materials must be clearly stated to guarantee ethical research conduct and consistent use of the resulting data.

Mitigating difficulties and adapting: the COVID-19 pandemic had a profound impact on the recruitment of patients with cancer between 2020 and 2022. However, centers continuously adapted their protocols to maintain patient inclusion rates and completeness of interview data and gained experience in new strategies for patient approach. For instance, the enrollment phase of patient identification, initial interview, information about the protocol, signature of informed written consent, lifestyle questionnaires, and some clinical information was conducted exclusively via telephone or video calls, and documents were provided electronically. These are strategies that were successfully piloted under these restrictive conditions and will continue to be used in other research protocols to facilitate patient recruitment and follow up.

Investing in large, geographically and population-diverse biorepositories: in the current largest international genomic consortium, PCAWG,[1] the overall ancestry distribution was heavily weighted toward donors of European descent (77% of total) followed by East Asians (16%), as expected for large contributions from European, North American, and Australian projects. Initial admixture analysis for four cancer sites in Mutographs shows that the percentages of donors of European, African, East Asian, and mixed-descent ancestry are, respectively, 25%, 30%, 32%, and 13% for ESCC; 90%, 1%, 5%, and 4% for RCC; 78%, 1%, 14%, and 7% for CRC; and 72%, 3%, 1%, and 24% for HNC. This reflects the geographical regions included based on incidence rates and risk exposures as previously discussed. Financial support is always required for participation in these large genomic epidemiology studies and should be allocated based on the contribution and needs of each collaborator. Funding bodies and cancer research agencies should envision additional financial and research capacity investment in large genomic epidemiology studies supporting participation of populations systematically underrepresented in this field.

Limitations of Mutographs and opportunities for future studies

There are cancer types not included in Mutographs that are highly relevant to understanding the etiology of cancer and should be considered in additional genomic epidemiology studies. For instance, high-incident cancers such as lung cancer adenocarcinoma, particularly prevalent in women, non-smokers, and Asian populations,[53] as well as less-incident, rare but frequently aggressive cancer types, for which many of the causes are still unknown. In addition, closer attention should be paid to the integration of genomics studies evaluating new exposures emerging as possible cancer risk factors, including air pollution,[54] vaping,[55] and opioid use.[56]

The knowledge generated by new large and geographically diverse biorepositories such as Mutographs have the potential to reveal previously unknown risk factors and ultimately establish causality, specifically by linking putative risk factors to specific genomic features. This, in turn, can guide the tailoring of prevention strategies and aid in the global reduction of the burden of cancer.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100500.

ACKNOWLEDGMENTS

This work was delivered as part of the Mutographs team supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (C98/A24032). Work at the Wellcome Sanger Institute was also supported by the Wellcome Trust (grants 206194 and 220540/Z/20/A), and work at the IARC/WHO was supported by regular budget funding, the NIH/NCI (grant number R21CA191965), and grant 2018/1795 from the Wereld Kanker Onderzoek Fonds (WKOF) as part of the World Cancer Research Fund International grant program. This work was supported in part by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. The head and neck cancer collection received funding from the European Union's Horizon 2020 research and innovation programme under grant no. 825771 and the São Paulo Research Foundation, FAPESP 2018/26297-3. Work at the Masaryk Memorial Cancer Institute, Brno, Czech Republic, was supported by MH CZ - DRO (MMCI, 00209805). The Porto Alegre center in Brazil received support from Hospital de Clínicas de Porto Alegre and Fundação Médica do Rio Grande do Sul. We are grateful for the support provided by the IARC General Services, including the Laboratory Services and Biobank team led by Z. Kozlakidis, the Section of Support to Research overseen by T. Landesz under IARC's regular budget funding, and the staff of DNA pipelines at the Wellcome Sanger Institute under the C98/A24032 grant. Farid Azmoudeh-Ardalan, Mojgan Asgari, Sophie Ferlicot, Hiva Saffar, Jean-Yves Scoazec, Stefano Serra, and Masoud Sotoudeh supported the pathological evaluation of samples. Laura Torrens Fontanals, Sergey Senkin, and Wellington Oliveira Dos Santos supported the admixture analysis. We are thankful for the work of all other collaborators in the Mutographs project who participated in the recruitment of patients in all centers. The authors would also like to thank all the patients and their families involved in these studies. We remember and celebrate Dr. Gloria Petersen for her passion for science, her inclusive intellect, and her kind and generous spirit.

Where authors are identified as personnel of the IARC/WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the IARC/WHO.

AUTHOR CONTRIBUTIONS

Conceptualization, S.P., B.A.-A., A.C.d.C., A.F.-I., and P.B.; funding acquisition, M.R.S., P.B., and L.B.A.; methodology, S.P., B.A.-A., A.C.d.C., A.F.-I., E.C., G.S., and P.B.; project administration, A.S. and L.H.; data curation, V.G., T. Cattiaux, and H.R.; resources, P.C., C. Carreira, R.C.C.P., A.N., S.O.A., P.A.-P., C. Canova, T. Chitapanaru, R.C., M.P.C., J.C.d.O., C.D., E.F., L. Ferri, R.F., L. Foretova, S.G., A.M.G., I.H., A.H., V.J., S.J., R.K., L.P.K., T.K., P.L., J.L., R.M., D. Mates, V.M., D. Menya, S. Mhatre, B.T.M., A.d.M., P.N., M.O., K.P., J.P., M.P.P., S.R., L.M.R.B., R.M.R., L.F.R.P., P.A.R.-U., S. Sangkhathat, S. Sangrajrang, T.S., E.S., B.S., C.V., J.R.V.d.P., N.S.V., M.V., J.Y., D.Z., K.Z., G.S., E.C., J.W., S.F., C.L., and S. Moody; supervision, S.P., B.A.-A., A.C.d.D., A.F.-I., and P.B.; visualization, S.P. and T. Cattiaux; writing – original draft, S.P., B.A.-A., A.C.d.D, A.F.-I., and P.B.; writing – review & editing, all authors.

DECLARATION OF INTERESTS

M.R.S. is founder of, consultant to, and stockholder in Quotient Therapeutics. L.B.A. is a compensated consultant and has equity interest in io9, LLC, and Genome Insight. His spouse is an employee of Biotheranostics, Inc. L.B.A.

is also an inventor of a US patent 10,776,718 for source identification by non-negative matrix factorization. L.B.A. declares US provisional applications with serial numbers 63/289,601; 63/269,033; and 63/483,237. L.B.A. also declares US provisional applications with serial numbers 63/366,392; 63/367,846; 63/412,835; and 63/492,348.

REFERENCES

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium; Abascal, F., Abeshouse, A., Aburatani, H., Adams, D.J., Agrawal, N., Ahn, K.S., Ahn, S.-M., Aikata, H., Akbani, R., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
2. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216. https://doi.org/10.1038/s41586-019-1689-y.
3. Degasperi, A., Zou, X., Amarante, T.D., Martinez-Martinez, A., Koh, G.C.C., Dias, J.M.L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V.Y.W., et al. (2022). Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. Science *376*, science.abl9283. https://doi.org/10.1126/science.abl9283.
4. Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A.R., and Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. Nat. Med. *28*, 243–250. https://doi.org/10.1038/s41591-021-01672-4.
5. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518. https://doi.org/10.1038/s41586-019-1310-4.
6. Spratt, D.E., Chan, T., Waldron, L., Speers, C., Feng, F.Y., Ogunwobi, O.O., and Osborne, J.R. (2016). Racial/Ethnic Disparities in Genomic Sequencing. JAMA Oncol. *2*, 1070–1074. https://doi.org/10.1001/jamaoncol.2016.1854.
7. Parkin, D.M. (1993). Studies of Cancer in Migrant Populations (IARC Sci Publ), pp. 1–10.
8. Piñeros, M., Laversanne, M., Barrios, E., Cancela, M.d.C., de Vries, E., Pardo, C., and Bray, F. (2022). An updated profile of the cancer burden, patterns and trends in Latin America and the Caribbean. Lancet Reg. Health. Am. *13*, 100294. https://doi.org/10.1016/j.lana.2022.100294.
9. Morgan, E., Soerjomataram, I., Rumgay, H., Coleman, H.G., Thrift, A.P., Vignat, J., Laversanne, M., Ferlay, J., and Arnold, M. (2022). The Global Landscape of Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma Incidence and Mortality in 2020 and Projections to 2040: New Estimates From GLOBOCAN 2020. Gastroenterology *163*, 649–658.e2. https://doi.org/10.1053/j.gastro.2022.05.054.
10. Znaor, A., Lortet-Tieulent, J., Laversanne, M., Jemal, A., and Bray, F. (2015). International variations and trends in renal cell carcinoma incidence and mortality. Eur. Urol. *67*, 519–530. https://doi.org/10.1016/j.eururo.2014.10.002.
11. Arnold, M., Abnet, C.C., Neale, R.E., Vignat, J., Giovannucci, E.L., McGlynn, K.A., and Bray, F. (2020). Global Burden of 5 Major Types of Gastrointestinal Cancer. Gastroenterology *159*, 335–349.e15. https://doi.org/10.1053/j.gastro.2020.02.068.
12. Taubes, G. (1995). Epidemiology faces its limits. Science *269*, 164–169. https://doi.org/10.1126/science.7618077.
13. Peto, J. (2001). Cancer epidemiology in the last century and the next decade. Nature *411*, 390–395. https://doi.org/10.1038/35077256.
14. McCullough, L.E., Maliniak, M.L., Amin, A.B., Baker, J.M., Baliashvili, D., Barberio, J., Barrera, C.M., Brown, C.A., Collin, L.J., Freedman, A.A., et al. (2022). Epidemiology beyond its limits. Sci. Adv. *8*, eabn3328. https://doi.org/10.1126/sciadv.abn3328.
15. Brennan, P., and Davey-Smith, G. (2022). Identifying Novel Causes of Cancers to Enhance Cancer Prevention: New Strategies Are Needed. J. Natl. Cancer Inst. *114*, 353–360. https://doi.org/10.1093/jnci/djab204.
16. Ogino, S., Nowak, J.A., Hamada, T., Milner, D.A., Jr., and Nishihara, R. (2019). Insights into Pathogenic Interactions Among Environment, Host, and Tumor at the Crossroads of Molecular Pathology and Epidemiology. Annu. Rev. Pathol. *14*, 83–103. https://doi.org/10.1146/annurev-pathmechdis-012418-012818.
17. Bray, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., R, Z., and Ferlay J, e. (2017). Cancer Incidence in Five Continents, *XI* (Lyon: International Agency for Research on Cancer), (electronic version).
18. (electronic version) Bray, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Zanetti, R., and Ferlay, J. (2017), Available from:. Cancer Incidence in Five Continents, *XI* (Lyon: International Agency for Research on Cancer), accessed on 22 March 2018. http://ci5.iarc.fr.
19. Arnold, M., Pandeya, N., Byrnes, G., Renehan, P.A.G., Stevens, G.A., Ezzati, P.M., Ferlay, J., Miranda, J.J., Romieu, I., Dikshit, R., et al. (2015). Global burden of cancer attributable to high body-mass index in 2012: a population-based study. Lancet Oncol. *16*, 36–46. https://doi.org/10.1016/S1470-2045(14)71123-4.
20. Bouvard, V., Loomis, D., Guyton, K.Z., Grosse, Y., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., Mattock, H., and Straif, K.; International Agency for Research on Cancer Monograph Working Group (2015). Carcinogenicity of consumption of red and processed meat. Lancet Oncol. *16*, 1599–1600. https://doi.org/10.1016/S1470-2045(15)00444-1.
21. Rumgay, H., Shield, K., Charvat, H., Ferrari, P., Sornpaisarn, B., Obot, I., Islami, F., Lemmens, V.E.P.P., Rehm, J., and Soerjomataram, I. (2021). Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. Lancet Oncol. *22*, 1071–1080. https://doi.org/10.1016/S1470-2045(21)00279-5.
22. IARC Monographs Vol 130 group (2021). Carcinogenicity of 1,1,1-trichloroethane and four other industrial chemicals. Lancet Oncol. *22*, 1661–1662. https://doi.org/10.1016/S1470-2045(21)00659-8.
23. Das, S., Thakur, S., Korenjak, M., Sidorenko, V.S., Chung, F.F.L., and Zavadil, J. (2022). Aristolochic acid-associated cancers: a public health risk in need of global action. Nat. Rev. Cancer *22*, 576–591. https://doi.org/10.1038/s41568-022-00494-x.
24. Grosse, Y., Baan, R., Straif, K., Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Galichet, L., and Cogliano, V.; WHO International Agency for Research on Cancer Monograph Working Group (2009). A review of human carcinogens–Part A: pharmaceuticals. Lancet Oncol. *10*, 13–14. https://doi.org/10.1016/s1470-2045(08)70286-9.
25. Scelo, G., Riazalhosseini, Y., Greger, L., Letourneau, L., Gonzàlez-Porta, M., Wozniak, M.B., Bourgey, M., Harnden, P., Egevad, L., Jackson, S.M., et al. (2014). Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat. Commun. *5*, 5135. https://doi.org/10.1038/ncomms6135.
26. Shearer, J.J., Callahan, C.L., Calafat, A.M., Huang, W.Y., Jones, R.R., Sabbisetti, V.S., Freedman, N.D., Sampson, J.N., Silverman, D.T., Purdue, M.P., and Hofmann, J.N. (2021). Serum Concentrations of Per- and Polyfluoroalkyl Substances and Risk of Renal Cell Carcinoma. J. Natl. Cancer Inst. *113*, 580–587. https://doi.org/10.1093/jnci/djaa143.
27. Alcala, K., Mariosa, D., Smith-Byrne, K., Nasrollahzadeh Nesheli, D., Carreras-Torres, R., Ardanaz Aicua, E., Bondonno, N.P., Bonet, C., Brunström, M., Bueno-de-Mesquita, B., et al. (2022). The relationship between blood pressure and risk of renal cell carcinoma. Int. J. Epidemiol. *51*, 1317–1327. https://doi.org/10.1093/ije/dyac042.
28. Weikert, S., Boeing, H., Pischon, T., Weikert, C., Olsen, A., Tjonneland, A., Overvad, K., Becker, N., Linseisen, J., Trichopoulou, A., et al. (2008). Blood pressure and risk of renal cell carcinoma in the European prospective investigation into cancer and nutrition. Am. J. Epidemiol. *167*, 438–446. https://doi.org/10.1093/aje/kwm321.
29. Ben, Q., Xu, M., Ning, X., Liu, J., Hong, S., Huang, W., Zhang, H., and Li, Z. (2011). Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of

cohort studies. Eur. J. Cancer *47*, 1928–1937. https://doi.org/10.1016/j.ejca.2011.03.003.

30. Ke, L. (2002). Mortality and incidence trends from esophagus cancer in selected geographic areas of China circa 1970-90. Int. J. Cancer *102*, 271–274. https://doi.org/10.1002/ijc.10706.
31. Abnet, C.C., Arnold, M., and Wei, W.Q. (2018). Epidemiology of Esophageal Squamous Cell Carcinoma. Gastroenterology *154*, 360–373. https://doi.org/10.1053/j.gastro.2017.08.023.
32. Sheikh, M., Poustchi, H., Pourshams, A., Etemadi, A., Islami, F., Khoshnia, M., Gharavi, A., Hashemian, M., Roshandel, G., Khademi, H., et al. (2019). Individual and Combined Effects of Environmental Risk Factors for Esophageal Cancer Based on Results From the Golestan Cohort Study. Gastroenterology *156*, 1416–1427. https://doi.org/10.1053/j.gastro.2018.12.024.
33. Mwachiro, M.M., Pritchett, N., Calafat, A.M., Parker, R.K., Lando, J.O., Murphy, G., Chepkwony, R., Burgert, S.L., Abnet, C.C., Topazian, M.D., et al. (2021). Indoor wood combustion, carcinogenic exposure and esophageal cancer in southwest Kenya. Environ. Int. *152*, 106485. https://doi.org/10.1016/j.envint.2021.106485.
34. Li, M., Park, J.Y., Sheikh, M., Kayamba, V., Rumgay, H., Jenab, M., Narh, C.T., Abedi-Ardekani, B., Morgan, E., de Martel, C., et al. (2023). Population-based investigation of common and deviating patterns of gastric cancer and oesophageal cancer incidence across populations and time. Gut *72*, 846–854. https://doi.org/10.1136/gutjnl-2022-328233.
35. Ryan, A.M., Duong, M., Healy, L., Ryan, S.A., Parekh, N., Reynolds, J.V., and Power, D.G. (2011). Obesity, metabolic syndrome and esophageal adenocarcinoma: epidemiology, etiology and new targets. Cancer Epidemiol. *35*, 309–319. https://doi.org/10.1016/j.canep.2011.03.001.
36. Macfarlane, T.V., Macfarlane, G.J., Oliver, R.J., Benhamou, S., Bouchardy, C., Ahrens, W., Pohlabeln, H., Lagiou, P., Lagiou, A., Castellsague, X., et al. (2010). The aetiology of upper aerodigestive tract cancers among young adults in Europe: the ARCAGE study. Cancer Causes Control. *21*, 2213–2221. https://doi.org/10.1007/s10552-010-9641-3.
37. Abrahão, R., Perdomo, S., Pinto, L.F.R., Nascimento de Carvalho, F., Dias, F.L., de Podestá, J.R.V., Ventorin von Zeidler, S., Marinho de Abreu, P., Vilensky, M., Giglio, R.E., et al. (2020). Predictors of Survival After Head and Neck Squamous Cell Carcinoma in South America: The InterCHANGE Study. JCO Glob. Oncol. *6*, 486–499. https://doi.org/10.1200/GO.20.00014.
38. Hadji, M., Rashidian, H., Marzban, M., Naghibzadeh-Tahami, A., Gholipour, M., Mohebbi, E., Safari-Faramani, R., Seyyedsalehi, M.S., Hosseini, B., Bakhshi, M., et al. (2022). Opium use and risk of bladder cancer: a multi-centre case-referent study in Iran. Int. J. Epidemiol. *51*, 830–838. https://doi.org/10.1093/ije/dyac031.
39. IARC Monographs Vol 126 group (2020). Carcinogenicity of opium consumption. Lancet Oncol. *21*, 1407–1408. https://doi.org/10.1016/S1470-2045(20)30611-2.
40. Matsuda, T., and Marugame, T. (2007). International comparisons of cumulative risk of gallbladder cancer and other biliary tract cancer, from Cancer Incidence in Five Continents Vol. VIII. Jpn. J. Clin. Oncol. *37*, 74–75. https://doi.org/10.1093/jjco/hyl158.
41. Maurice, J. (2016). IARC celebrates 50 years of cancer research. Lancet *387*, 2367–2368. https://doi.org/10.1016/S0140-6736(16)30784-X.
42. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J.G. (2009). Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. *42*, 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.
43. Moody, S., Senkin, S., Islam, S.M.A., Wang, J., Nasrollahzadeh, D., Cortez Cardoso Penha, R., Fitzgerald, S., Bergstrom, E.N., Atkins, J., He, Y., et al. (2021). Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat. Genet. *53*, 1553–1563. https://doi.org/10.1038/s41588-021-00928-6.
44. Senkin, S., Moody, S., Díaz-Gay, M., Abedi-Ardekani, B., Cattiaux, T., Ferreiro-Iglesias, A., Wang, J., Fitzgerald, S., Kazachkova, M., Vangara, R., et al. (2023). Geographic variation of mutagenic exposures in kidney cancer genomes. Preprint at medRxiv *888*. https://doi.org/10.1101/2023.06.20.23291538.
45. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.
46. International Cancer Genome Consortium, ICGC ARGO (2022). http://platform.icgc-argo.org/.
47. Koh, G., Degasperi, A., Zou, X., Momen, S., and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. Nat. Rev. Cancer *21*, 619–637. https://doi.org/10.1038/s41568-021-00377-7.
48. Mustjoki, S., and Young, N.S. (2021). Somatic Mutations in "Benign" Disease. N. Engl. J. Med. *384*, 2039–2052. https://doi.org/10.1056/NEJMra2101920.
49. Balmain, A. (2020). The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk. Nat. Genet. *52*, 1139–1143. https://doi.org/10.1038/s41588-020-00727-5.
50. Kakiuchi, N., and Ogawa, S. (2021). Clonal expansion in non-cancer tissues. Nat. Rev. Cancer *21*, 239–256. https://doi.org/10.1038/s41568-021-00335-3.
51. Wijewardhane, N., Dressler, L., and Ciccarelli, F.D. (2021). Normal Somatic Mutations in Cancer Transformation. Cancer Cell *39*, 125–129. https://doi.org/10.1016/j.ccell.2020.11.002.
52. Colom, B., Herms, A., Hall, M.W.J., Dentro, S.C., King, C., Sood, R.K., Alcolea, M.P., Piedrafita, G., Fernandez-Antoran, D., Ong, S.H., et al. (2021). Mutant clones in normal epithelium outcompete and eliminate emerging tumours. Nature *598*, 510–514. https://doi.org/10.1038/s41586-021-03965-7.
53. Zhang, Y., Vaccarella, S., Morgan, E., Li, M., Etxeberria, J., Chokunonga, E., Manraj, S.S., Kamate, B., Omonisi, A., and Bray, F. (2023). Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. Lancet Oncol. *24*, 1206–1218. https://doi.org/10.1016/S1470-2045(23)00444-8.
54. Hill, W., Lim, E.L., Weeden, C.E., Lee, C., Augustine, M., Chen, K., Kuan, F.-C., Marongiu, F., Evans, E.J., Moore, D.A., et al. (2023). Lung adenocarcinoma promotion by air pollutants. Nature *616*, 159–167. https://doi.org/10.1038/s41586-023-05874-3.
55. Goniewicz, M.L., Smith, D.M., Edwards, K.C., Blount, B.C., Caldwell, K.L., Feng, J., Wang, L., Christensen, C., Ambrose, B., Borek, N., et al. (2018). Comparison of Nicotine and Toxicant Exposure in Users of Electronic Cigarettes and Combustible Cigarettes. JAMA Netw. Open *1*, e185937. https://doi.org/10.1001/jamanetworkopen.2018.5937.
56. Sheikh, M., Brennan, P., Mariosa, D., and Robbins, H.A. (2023). Opioid medications: an emerging cancer risk factor? Br. J. Anaesth. *130*, e401–e403. https://doi.org/10.1016/j.bja.2022.12.007.

Cell Genomics, Volume *4*

Supplemental information

The Mutographs biorepository:

A unique genomic resource

to study cancer around the world

Sandra Perdomo, Behnoush Abedi-Ardekani, Ana Carolina de Carvalho, Aida Ferreiro-Iglesias, Valérie Gaborieau, Thomas Cattiaux, Hélène Renard, Priscilia Chopard, Christine Carreira, Andreea Spanu, Arash Nikmanesh, Ricardo Cortez Cardoso Penha, Samuel O. Antwi, Patricia Ashton-Prolla, Cristina Canova, Taned Chitapanarux, Riley Cox, Maria Paula Curado, José Carlos de Oliveira, Charles Dzamalala, Elenora Fabianova, Lorenzo Ferri, Rebecca Fitzgerald, Lenka Foretova, Steven Gallinger, Alisa M. Goldstein, Ivana Holcatova, Antonio Huertas, Vladimir Janout, Sonata Jarmalaite, Radka Kaneva, Luiz Paulo Kowalski, Tomislav Kulis, Pagona Lagiou, Jolanta Lissowska, Reza Malekzadeh, Dana Mates, Valerie McCormack, Diana Menya, Sharayu Mhatre, Blandina Theophil Mmbaga, André de Moricz, Péter Nyirády, Miodrag Ognjanovic, Kyriaki Papadopoulou, Jerry Polesel, Mark P. Purdue, Stefan Rascu, Lidia Maria Rebolho Batista, Rui Manuel Reis, Luis Felipe Ribeiro Pinto, Paula A. Rodríguez-Urrego, Surasak Sangkhathat, Suleeporn Sangrajrang, Tatsuhiro Shibata, Eduard Stakhovsky, Beata Świątkowska, Carlos Vaccaro, Jose Roberto Vasconcelos de Podesta, Naveen S. Vasudev, Marta Vilensky, Jonathan Yeung, David Zaridze, Kazem Zendehdel, Ghislaine Scelo, Estelle Chanudet, Jingwei Wang, Stephen Fitzgerald, Calli Latimer, Sarah Moody, Laura Humphreys, Ludmil B. Alexandrov, Michael R. Stratton, and Paul Brennan

Cell Genomics Transparent Peer Review Record

The Mutographs biorepository: A unique genomic resource to study cancer around the world

Sandra Perdomo[1‡§], Behnoush Abedi-Ardekani[1‡], Ana Carolina de Carvalho[1‡], Aida Ferreiro-Iglesias[1‡], Valérie Gaborieau[1], Thomas Cattiaux[1], Hélène Renard[1], Priscilia Chopard[1], Christine Carreira[2], Andreea Spanu[1], Arash Nikmanesh[1], Ricardo Cortez Cardoso Penha[1], Samuel O. Antwi[3,4], Patricia Ashton-Prolla[5,6], Cristina Canova[7], Taned Chitapanarux[8], Riley Cox[9], Maria Paula Curado[10], José Carlos de Oliveira[11], Charles Dzamalala[12], Elenora Fabianova[13], Lorenzo Ferri[14], Rebecca Fitzgerald[15], Lenka Foretova[16], Steven Gallinger[17], Alisa M. Goldstein[18], Ivana Holcatova[19,20], Antonio Huertas[21], Vladimir Janout[22], Sonata Jarmalaite[23,24], Radka Kaneva[25], Luiz Paulo Kowalski[10,26], Tomislav Kulis[27,28], Pagona Lagiou[29], Jolanta Lissowska[30], Reza Malekzadeh[31], Dana Mates[32], Valerie McCormack[33], Diana Menya[34], Sharayu Mhatre[35], Blandina Theophil Mmbaga[36], André de Moricz[37], Péter Nyirády[38], Miodrag Ognjanovic[39], Kyriaki Papadopoulou[40], Jerry Polesel[41], Mark P. Purdue[42], Stefan Rascu[43], Lidia Maria Rebolho Batista[44], Rui Manuel Reis[44,45], Luis Felipe Ribeiro Pinto[46], Paula A. Rodríguez-Urrego[47], Surasak Sangkhathat[48], Suleeporn Sangrajrang[49], Tatsuhiro Shibata[50,51], Eduard Stakhovsky[52], Beata Świątkowska[53], Carlos Vaccaro[54], Jose Roberto Vasconcelos de Podesta[55], Naveen S. Vasudev[56], Marta Vilensky[57], Jonathan Yeung[58], David Zaridze[59], Kazem Zendehdel[60], Ghislaine Scelo[61], Estelle Chanudet[62], Jingwei Wang[63], Stephen Fitzgerald[63], Calli Latimer[63], Sarah Moody[63], Laura Humphreys[63], Ludmil B. Alexandrov[64,65,66], Michael R. Stratton[63], Paul Brennan[1]*. On behalf of the Mutographs Study

Cell Genomics Transparent Peer Review Record

Summary

Initial submission:	Received : 7/17/2023
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 8/23/2023 Revision received : 10/24/2023
Second round of review:	Number of reviewers: 2 Accepted : 1/10/2023
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: The article focuses on a description of the Mutographs Grand Challenge, funded by Cancer Research UK. The Mutographs project aims at collecting thousands of cancer whole genomes, as well as patient information, in 30 countries across the world. This is an important project and first one of its kind, which will with no doubt improve our understanding of the mutagenic and environmental causes of cancer. Other cancer sequencing efforts are often limited to one or very few countries involved, and the patient information tends to be incomplete.

Mutographs tries to overcome these limitations by coordinating sequencing and efforts and patient data collection and harmonisation across several countries.

While the article reads well and does a good job to motivate the Mutographs effort, most of the article, especially from line 138, feels to me like a long materials and methods section. I believe that the article would benefit from reducing the description of how the samples and the data are collected, and from adding more information pertinent to the Perspective style. Some examples could be more description and details of past and current efforts and their findings and limitations, or looking ahead perhaps more at the potential impacts or at need and the obstacles to bring efforts like the Mutographs to even more countries, even those that lag behind in technology and investments.

In conclusion, I believe the article requires some work to make it a more informative and enjoyable read.

Reviewer #2: This project presents an unprecedented data source for cancer research, offering a comprehensive collection of phenotypic information across various research centers. However, there is room for improvement especially in genetic data generation and processing. Here are my comments.
1. On page 3, regarding the questions asked by the journal, I believe the rationales of both questions are "Yes". It is crucial to provide the analysis code for Whole Genome Sequencing (WGS) results, including, at a minimum, information about the software version and parameters employed. It's also noteworthy that this project aims to generate a substantial, large-scale dataset.
2. How does this dataset correlate with existing datasets in ICGC or TCGA? Is there any overlap, particularly since a significant portion of the data were from aggregating data from pre-existing sources like data pools and biorepositories? Could this imply an overlap with previous patient cases? Furthermore, could you clarify the concept of pooling data from existing sources and biorepositories? Does this signify that cases were sourced from established databases or biobanks? A verification of genetic relatedness would be essential.
3. From lines 217 to 222, there are some abbreviations in the text that do not appear in Table 1, and vice versa. For enhanced clarity, aligning the information consistently would be beneficial.
4. In the same section (lines 217-222), why are only these three cancer sites mentioned? Does this imply that PCAWG lacks exposure information for other cancer types? Is there a specific rationale behind highlighting the place of residence?
5. Regarding lines 255-256, it's imperative to describe the sequencing protocol employed. If the protocol aligns with reference 38, it's important to acknowledge that the sequencing depth is notably lower than that of PCAWG (https://www.nature.com/articles/s41586-020-1969-6#Sec14). Additionally, differences in sequence read length need clarification. Providing a concise overview of the data analysis pipeline, including software and version details, is advisable.
6. Within Table 1, the presence of samples with unknown gender raises questions. Given that gender can be determined from genetic data, could you elucidate the reason behind this discrepancy?

Authors' response to the first round of review

We thank the reviewers and the editor for your valuable time and for the enriching feedback that undoubtedly will improve the quality of this article titled "The Mutographs biorepository: A unique genomic resource to study cancer around the world". In response to the reviewers' and Editor's comments, we have made the following changes to the manuscript:

Reviewers' Comments:
Reviewer #1:
While the article reads well and does a good job to motivate the Mutographs effort, most of the article, especially from line 138, feels to me like a long materials and methods section. I believe that the article would benefit from reducing the description of how the samples and the data are collected, and from adding more information pertinent to the

Perspective style. Some examples could be more description and details of past and current efforts and their findings and limitations, or looking ahead perhaps more at the potential impacts or at need and the obstacles to bring efforts like the Mutographs to even more countries, even those that lag behind in technology and investments.
In conclusion, I believe the article requires some work to make it a more informative and enjoyable read.

We agree with the reviewer's suggestion, and we restructured the manuscript to include more information that could emphasize the uniqueness of the Mutographs study collection, the opportunities to extend and/or develop similar initiatives in other countries and the lessons learned during the development of this biorepository.

-We have included a new section explaining the overall rationale of the Mutographs study and more details explaining why this is an example of how genomic epidemiology studies in comparison to classical epidemiological studies constitute better approaches to identify new causes of cancer globally (lines 34-53).
-We included additional examples (line 332) of future uses of the samples and analyses of the data generated by the Mutographs study.
-A final section (lines 381-456) highlights some of the fundamental aspects that contributed to the creation of this large-scale cancer biorepository and that could be used in future similar initiatives

Reviewer #2:
1. On page 3, regarding the questions asked by the journal, I believe the rationales of both questions are "Yes". It is crucial to provide the analysis code for Whole Genome Sequencing (WGS) results, including, at a minimum, information about the software version and parameters employed. It's also noteworthy that this project aims to generate a substantial, large-scale dataset.

We have included additional information as part of the Data repository and sharing section (lines 307-327). We incorporated the study reference for the data submitted to EGA, references to the sequencing pipelines already published and the links to bioinformatic repositories which include the algorithms, software versions and codes used in the analyses.

2. How does this dataset correlate with existing datasets in ICGC or TCGA? Is there any overlap, particularly since a significant portion of the data were from aggregating data from pre-existing sources like data pools and biorepositories? Could this imply an overlap with previous patient cases? Furthermore, could you clarify the concept of pooling data from existing sources and biorepositories? Does this signify that cases were sourced from established databases or biobanks? A verification of genetic relatedness would be essential.

We have clarified this point in the description of methods (lines 159-162). The retrospective studies and biorepository collections contributing to Mutographs have been selected from among those that have not been previously analysed or included in other large international genomic projects, mainly ICGC and TCGA. Therefore, the sequencing data and metadata contributing to Mutographs are not integrated into any other publicly available dataset. Similarly, samples selected in Mutographs from existing biorepositories have never been selected for sequencing in previous genomics initiatives.

In the description of data collection, we refer to harmonising rather to pooling data (line 226). For instance, data on smoking\alcohol history from existing biorepositories was harmonised to evaluate history of exposure, quantity, and frequency using the same definition for each variable. This has been described in the published data from Mutographs (references added in the corresponding section).

We have modified the sentence in line 124. We removed the term "pooling data" to avoid confusion.

3. From lines 217 to 222, there are some abbreviations in the text that do not appear in Table 1, and vice versa. For enhanced clarity, aligning the information consistently would be beneficial.

We have included the abbreviations in table 1 and in the table legend to keep consistency with the text.

4. In the same section (lines 217-222), why are only these three cancer sites mentioned? Does this imply that PCAWG lacks exposure information for other cancer types? Is there a specific rationale behind highlighting the place of residence?

We compared the available exposure information among patients from cancer sites included in both PCAWG and Mutographs to estimate the extent of the Mutographs project metadata. These three cancer sites: Esophageal, Head and Neck and Pancreas were the only overlapping cancer sites between the two studies with exposure information. PCAWG only included partial information on history of alcohol and tobacco consumption as highlighted in the text (lines 209-216) and in Table 1. The updated version of Table 1 includes the countries from where cases were selected in both studies to emphasise a higher geographical diversity in our study.

Information regarding residential history was important to highlight specific lifestyle and/or environmental and risk exposures for different regions, i.e consumption of opium for specific regions in the north of Iran and mate drinking in the south of Brazil. To understand exposure to Aristolochic acid in the Balkan region and boundary countries, we intended to use residential history to track a possible environmental source of exposure to this carcinogen beyond the use of herbal remedies.

5. Regarding lines 255-256, it's imperative to describe the sequencing protocol employed. If the protocol aligns with reference 38, it's important to acknowledge that the sequencing depth is notably lower than that of PCAWG (https://www.nature.com/articles/s41586-020-1969-6#Sec14). Additionally, differences in sequence read length need clarification. Providing a concise overview of the data analysis pipeline, including software and version details, is advisable.

We have added specific details on the depth of coverage for both tumor (40X) and normal tissues (20X)(lines 282-287) and minimal depth considered for further analyses. As mentioned by the reviewer, in PCAWG, the mean read coverage was 39X (higher than in our study) for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38X and 60X (within the range for tumors sequenced in Mutographs).
We also added the references to the sequencing pipelines from published articles and the links to bioinformatic repositories which include all algorithms, software versions and bioinformatic codes used in the analyses as clarified above in point 1.

6. Within Table 1, the presence of samples with unknown gender raises questions. Given that gender can be determined from genetic data, could you elucidate the reason behind this discrepancy?

In the new version of Table 1. we have completed the missing information on sex from cases in Mutographs. However, metadata publicly available and published for the PDAC cases in the PCAWG collection lacked information on sex for 2 cases. https://doi.org/10.1038/s41586-020-1969-6. Supplementary Table 1.

Referees' reports, second round of review

Reviewer #1: I would like to thank the authors for working to address my previous comments. I am satisfied with the changes, and I believe that the article is now much improved. The authors provide a clear motivation for the Mutographs projects, contextualise it and provide a discussion of their workflow and how they addressed the challenges of the project. Finally, they illustrate how their initiative can serve as an example and pave the way for other projects that can build on it.

My only remaining concern is that the project is still ongoing and, while one article and its data have been published in 2021 (Moody et al, about 552 esophageal cancers), most of the data and

the results from the Mutographs project as it has been described here (4,400 successfully processed samples) are not yet available/published. This limits to some extent the discussion of the impact and effects that this project has had so far. At the same time, I think that it might be up to the editor to decide whether we should wait or not for more results to be published, so that a summary of the results and their impact can be included. The presented article certainly has already the potential to be a very influential piece, illustrating how efforts of cancer sequencing involving many countries globally with highly harmonised metadata are possible and still very much needed.

Reviewer #2: 1. The line numbers in authors' responses do not align with the resubmitted revision. For example, the author indicated the depth of coverage contents were in line 282-278, but it actually are in line 248-250. Please make sure line numbers are updated to correspond with the latest revision of the manuscript.

2. If the line numbers are correct, there appears to be a discrepancy between lines 124-125 and the response to question 2 from reviewer 2 concerning the term "pooling data." If this phrase does not accurately represent your intended meaning, please amend the text in lines 124-125 to ensure consistency across the manuscript.

3. The author did not modify the color of Figure 1 accordingly.

Authors' response to the second round of review

We thank the reviewers and the editor for their final comments and remarks. In response to those comments, we have made the following changes to the manuscript:

Reviewer #1:

I would like to thank the authors for working to address my previous comments. I am satisfied with the changes, and I believe that the article is now much improved. The authors provide a clear motivation for the Mutographs projects, contextualise it and provide a discussion of their workflow and how they addressed the challenges of the project. Finally, they illustrate how their initiative can serve as an example and pave the way for other projects that can build on it.

The comments from the reviewer were extremely useful to reshape the focus of the manuscript. We emphasized the uniqueness of the Mutographs study collection, the opportunities to extend and/or develop similar initiatives in other countries and the lessons learned during the development of this biorepository.

My only remaining concern is that the project is still ongoing and, while one article and its data have been published in 2021 (Moody et al, about 552 esophageal cancers), most of the data and the results from the Mutographs project as it has been described here (4,400 successfully processed samples) are not yet available/published. This limits to some extent the discussion of the impact and effects that this project has had so far. At the same time, I think that it might be up to the editor to decide whether we should wait or not for more results to be published, so that a summary of the results and their impact can be included. The presented article certainly has already the potential to be a very influential piece, illustrating how efforts of cancer sequencing involving many countries globally with highly harmonised metadata are possible and still very much needed.

We acknowledge that the analyses included in the Mutographs

project are still ongoing. The end of the project was extended until January 2025. However, a great progress has been accomplished until now. For instance, the kidney cancer analysis (in biorchives REF43) is now under review after resubmission to Nature. New publications will follow in 2024, two manuscripts are currently in preparation, the Head and Neck cancer analysis to be submited in mid January and the Colorectal cancer manuscript in Spring 2024. We envision that the publication of the Mutographs Biorepository will enhance visibility to the new upcoming publications and vice versa.
As suggested by the editor we also included a paragraph on Limitations of the Study in the discussion section and added new supporting references.

Response to Reviewers
Reviewer #2:
3. The author did not modify the color of Figure 1 accordingly.
The Map in Figure 1B (Now labeled Figure 2) has been updated accordingly. For clarity, we decided to color in blue all the countries included in Mutographs and point to the cities included in the patient collection. The complete list of cities per country is included in the corresponding figure legend.

2. If the line numbers are correct, there appears to be a discrepancy between lines 124-125 and the response to question 2 from reviewer 2 concerning the term "pooling data." If this phrase does not accurately represent your intended meaning, please amend the text in lines 124-125 to ensure consistency across the manuscript.

We have modified the sentence in line 124. We removed the term "pooling data" to avoid confusion.

1. The line numbers in authors' responses do not align with the resubmitted revision. For example, the author indicated the depth of coverage contents were in line 282-278, but it actually are in line 248-250. Please make sure line numbers are updated to correspond with the latest revision of the manuscript.

We have added specific details on the depth of coverage for both tumor (40X) and normal tissues (20X)(lines 246-250) and minimal depth considered for further analyses. As mentioned by the reviewer, in PCAWG, the mean read coverage was 39X (higher than in our study) for normal samples, whereas tumours had a bimodal coverage distribution with modes at 38X and 60X (within the range for tumors sequenced in Mutographs).