BMJ Open

GCAT | Genomes for life: a prospective cohort study of the genomes of Catalonia

Mireia Obón-Santacana,[1,2] Mireia Vilardell,[1] Anna Carreras,[1] Xavier Duran,[1] Juan Velasco,[1] Iván Galván-Femenía,[1] Teresa Alonso,[1] Lluís Puig,[3] Lauro Sumoy,[4] Eric J Duell,[5] Manuel Perucho,[4] Victor Moreno,[2,6,7] Rafael de Cid[1]

To cite: Obón-Santacana M, Vilardell M, Carreras A, *et al*. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* 2018;**8**:e018324. doi:10.1136/bmjopen-2017-018324

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-018324).

Received 20 June 2017
Revised 3 January 2018
Accepted 1 February 2018

For numbered affiliations see end of article.

Correspondence to
Dr Rafael de Cid;
Rdecid@igtp.cat

ABSTRACT

Purpose The prevalence of chronic non-communicable diseases (NCDs) is increasing worldwide. NCDs are the leading cause of both morbidity and mortality, and it is estimated that by 2030, they will be responsible for 80% of deaths across the world. The Genomes for Life (GCAT) project is a long-term prospective cohort study that was designed to integrate and assess the role of epidemiological, genomic and epigenomic factors in the development of major chronic diseases in Catalonia, a north-east region of Spain.

Participants At the end of 2017, the GCAT Study will have recruited 20 000 participants aged 40–65 years. Participants who agreed to take part in the study completed a self-administered computer-driven questionnaire, and underwent blood pressure, cardiac frequency and anthropometry measurements. For each participant, blood plasma, blood serum and white blood cells are collected at baseline. The GCAT Study has access to the electronic health records of the Catalan Public Healthcare System. Participants will be followed biannually at least 20 years after recruitment.

Findings to date Among all GCAT participants, 59.2% are women and 83.3% of the cohort identified themselves as Caucasian/white. More than half of the participants have higher education levels, 72.2% are current workers and 42.1% are classified as overweight (body mass index ≥25 and <30 kg/m^2). We have genotyped 5459 participants, of which 5000 have metabolome data. Further, the whole genome of 808 participants will be sequenced by the end of 2017.

Future plans The first follow-up study started in December 2017 and will end by March 2018. Residences of all subjects will be geocoded during the following year. Several genomic analyses are ongoing, and metabolomic and genomic integrations will be performed to identify underlying genetic variants, as well as environmental factors that influence metabolites.

Strengths and limitations of this study

- The only currently available population-based prospective cohort study in Spain with more than 5500 genotyped participants and 800 whole genome sequences.
- Long period of follow-up: 20 years; participants will be contacted again biannually.
- Blood plasma, blood serum and white blood cells are collected and stored at baseline for each participant; detailed epidemiological and anthropometric measurements; access to electronic health records (EHR)of the Catalan Public Healthcare System that will allow researchers to have both retrospective and prospective data.
- The Genomes for Life (GCAT) Study will integrate and assess the role of epidemiological, environmental, EHR and omic factors in the development of chronic diseases (ie, by using molecular pathological epidemiology analyses).
- The GCAT cohort is mainly based on volunteer members of the Blood and Tissue Bank of Catalonia.

INTRODUCTION

The prevalence of chronic non-communicable diseases (NCDs) is increasing worldwide.[1–3] NCDs, such as cardiovascular diseases, cancer, respiratory diseases and diabetes, are characterised as having a long duration and a slow progression. NCDs are currently the leading cause of both morbidity and mortality, and it is estimated that by 2030, they will be responsible for 80% of deaths across the world.[4] Cancer affected around 3.45 million Europeans in 2012 and caused 1.75 million deaths.[5] In 2015, almost 248 000 new cancer cases were diagnosed in Spain, with colorectal, prostate, lung, breast and urinary bladder cancers being the five most common.[6] The morbidity and mortality rates of these conditions, together with other chronic disorders (ie, obesity, asthma, arthritis) are responsible for the high burden on public healthcare system expenses. Therefore, there exists a huge interest in developing and implementing new predictive and prognostic methods, as well as in adopting new public health strategies to reduce their socioeconomic impact.[7]

Individual susceptibilities to develop NCDs as well as their progression are influenced by genetic, epigenetic and environmental factors, and their interaction (also known as gene-environment interaction).[8–10] During the last decade, many studies have assessed the association between genetic variability and disease, both with candidate gene approaches and with comprehensive and agnostic genome-wide analysis (genome-wide association studies; GWAS). These studies have focused on common variant analysis of single nucleotide polymorphisms (SNPs) and have identified more than 36 000 risk loci for more than 60 common diseases.[11] However, the relative risks (RRs) reported are too low to be clinically relevant, and do not take into consideration the contribution of rare and structural variants.[12 13] There is strong evidence that rare genetic variation is important for disease predisposition.[14 15] Next-generation sequencing technologies allow the identification of novel rare variants, and may aid in increasing our understanding of the biology of cancer susceptibility and complex traits.[16] The role of epigenetic variation in disease susceptibility is an important factor to consider, either influenced by underlying genetic variants or modulated by the impact of the environment.

There is also robust epidemiological evidence from ecological studies that changes in the environmental exposures affect cancer and other diseases' incidence and mortality, suggesting that genetic predisposition cannot explain the whole incidence/mortality variability between countries.[17–19] In fact, WHO listed tobacco, high blood pressure, overweight and obesity, physical inactivity, high blood glucose, high cholesterol, low fruit and vegetable intake, urban outdoor air pollution, alcohol consumption, and occupational risks as the major risk factors in high-income countries.[20] Thus, it is important to design and develop new primary and secondary prevention strategies to reduce these exposures or mitigate their impact on NCD incidence and mortality.

Cohort studies have long been used to study determinants of disease and are considered to produce the highest level of evidence among observational studies. Longitudinal studies, such as prospective cohort studies, have a straightforward design. Information is collected at the time of recruitment, when the study population is free of disease, preferentially with repeated measurements. The population is followed over time, until the emergence of the outcomes of concern (ie, cancer, cardiovascular disease, diabetes, asthma). This design allows researchers to evaluate comparisons between exposed and unexposed subjects, and assess the magnitude of the associations using relative and absolute measures of risk or effect, and are less to prone to information biases than retrospective designs.[21]

The Genomes for Life (GCAT) Study is a long-term project that was set up to integrate and assess the role of epidemiological, environmental and omic factors (ie, genomic, metabolomic, proteomic, epigenomic) in the development of chronic diseases. Furthermore, GCAT also aims to assess the prevalence of risk factors and their association with disease incidence over time. Different but complementary lines of research will be pursued between genetic susceptibility and potential risk factors (the large sample size for some NCDs will allow the study of gene-gene and gene-environment interactions), the relation between several biomarkers in blood (ie, dietary, inflammatory, metabolomic, hormonal) and diseases, and the associations between epidemiological risk factors and diseases. These objectives will provide an exceptional opportunity to explore the association between the genome and the phenome of a large number of participants, since the GCAT Study will address different outcomes. The present article provides a comprehensive description of the GCAT Study.

COHORT DESCRIPTION

Study design, population, recruitment

The GCAT project is a prospective cohort study that was designed to recruit the general population of the north-east region of Spain, Catalonia, with a population of 7522 596 inhabitants. From April 2014 to June 2014, a pilot study (including 191 participants) was conducted in two centres to assess the feasibility of the study, and thereafter the project started.

The cohort is open to any volunteer that requests to participate; however, to improve recruitment, the GCAT cohort are individuals mostly enrolled from blood donors invited through the Blood and Tissue Bank (BST), a public agency of the Catalan Department of Health that guarantees the supply and proper use of human blood and tissue in Catalonia (http://www.bancsang.net/en_index/). With the aim to identify chronic disease events in the mid-term, the study covers a middle-aged range (40–65 years old) corresponding to 30% of the Catalan population.[22] In addition, participants are required to be able to understand at least one of the two official languages in Catalonia (Catalan or Spanish) to provide written informed consent, to possess an Individual Health System Identification Card and to be current residents of Catalonia. Potential participants are excluded if they have mental or health impairment disorders that impede giving written informed consent or efficient communication, or if they are planning to leave Catalonia during the following 5 years.

Participants are invited to participate using multiple active strategies, such as phone call, mail, GCAT web page (http://www.genomesforlife.com/participants/) or in person. Then, an appointment is agreed on and participants are asked to attend a recruitment centre. There are 11 permanent recruitment centres (figure 1).

Although there is no attempt to obtain a truly representative sample of the general population, in addition to the permanent centres, a large number of temporal recruitment centres are been organised all over Catalonia to accelerate recruitment. Recruitment is also open to any volunteer who meets the above criteria and is willing to participate. In this case, volunteers should ask for an

Figure 1 Genomes for Life (GCAT) recruitment centres. Distribution of the 11 GCAT permanent recruitment centres across the Catalan territory.

appointment by phone or via our GCAT web page after filling in a registration form. All participants who agree to be part of the study provided an informed consent and are asked to sign a consent agreement form that allows permission to access electronic health records (EHRs) for passive follow-up and to be contacted regularly to collect follow-up information collection on lifestyle and disease events. Participants are free to leave the study or withdraw their consent for specific areas of research.

Participants who agree to take part in the study complete an epidemiological questionnaire, donate a blood sample and undergo blood pressure, cardiac frequency and anthropometry measurements. All biological and physical examinations are performed in a separate room. Baseline interviews are performed by trained healthcare professionals (doctors and nurses). Specific guidelines were designed by the GCAT scientific members to support the interviewers, and to ensure uniform data collection.

Epidemiological questionnaire

Epidemiological interviews are done in a designated area using dedicated computers to ensure privacy. The electronic computer-based epidemiological questionnaire is included in the *eGCAT* software, which allows a comprehensive tracking of all the recruitment process.[23] The *eGCAT* is an adapted version of Onyx (www.obiba.org).[24] Customisation in local languages was performed in collaboration with the software developers at the Maelstrom Research Group, Research Institute of the McGill University Health Centre Montreal, Canada. A paper questionnaire was also designed in case of system failure or computer illiteracy.

Participants complete a self-administered computer-based questionnaire that collects data on a large number of lifestyle and health factors that are of interest in epidemiological and genetic studies. The GCAT baseline epidemiological questionnaire was specially designed to facilitate interoperability and collaboration with other survey studies. All variables measured in the GCAT Study are grouped in 'Group Theme and Domain', as proposed by the international guidelines for harmonisation of prospective population-based cohorts.[24 25]

The baseline survey includes 142 and 149 questions for men and women, respectively. Detailed information is also assessed on sociodemographic and socioeconomic status, current and past occupation, physical activity, lifetime tobacco and alcohol consumption, diet, personal and familiar medical history (parents, sisters/brothers and sons/daughters), prescription drug use, as well as specific questions related to women's or men's health. All epidemiological variables can be examined at the MICA repository, a web application used to create web data portals for epidemiological or consortium studies (http://gattaca.imppc.org/gcat-mica/mica/study/gcat).

Sociodemographic, socioeconomic and occupational variables assessment

Participants are required to fill in information on their gender, date and country of birth, current residence, ethnicity, laterality, marital status, social network, household incomes and type of healthcare access. Education levels are categorised as low (primary school, none), middle (vocational, secondary school, high school) and high (vocational postsecondary school, university studies or equivalent).

At enrolment, participants are asked about their current occupational status and type of job. The occupations asked are categorised based on the Spanish National Occupation Classification (CNO-11),[26] which derives from the International Standard Classification of Occupations (ISCO-08).[27] For each job reported, detailed questions to ascertain time schedules (rotating, morning work, evening work, split duty, night work), total hours worked per week and occupational physical activity are assessed.

Tobacco and alcohol assessment

Detailed questions on lifetime history of tobacco smoking (including cigarettes, cigars, pipe, hand-rolling tobacco, electronic cigarettes and waterpipe tobacco) address information on current status, smoking intensity, total lifetime dosage of tobacco smoke (measured in pack-years), age at initiation and cessation, and current/former number of cigarettes smoked per day. Further, second-hand smoke exposure at home and at work are assessed both during childhood and adulthood. An adapted and reduced version of the Fagerström Test, also known as the Heaviness of Smoking Index, is used to estimate nicotine dependence.[28]

Participants report the average number of standard glasses of wine, beer, champagne, sweet liquor or distilled spirits drunk per day or per week over the year before recruitment. Average volume of alcohol consumption per day is assessed using the 'standard drink unit', which is equivalent to 10 g of ethanol, and has been validated and extensively used in Spanish cohorts.[29] Participants are categorised based on drinking status, drinking patterns or alcohol intake following the WHO-alcohol classification.[30] Gender differences in drinking habits are also taken into account:[31] men are classified into six categories (former, never, low, moderate, high and very high intake) whereas five categories are defined for women (former, never, low, moderate and high intake).

Physical activity assessment

A validated short version of the European Prospective Investigation into Cancer and Nutrition (EPIC) Physical Activity Questionnaire (PAQ) is used to assess free-living activity referring to the past 12 months.[32] The GCAT Questionnaire slightly differed from the PAQ version, since the most frequent sports in Catalonia/Spain were asked. All physical activities are coded using the Compendium of Physical Activities, and metabolic equivalent hours per week value are used to denote intensity.[33] Following the recommendations, the 'total physical activity index' is assessed in three domains (leisure, occupational and housework PA), so that participants can be categorised into four levels (inactive, moderately inactive, moderately active and active).[34]

Dietary assessment

To estimate baseline adherence to the Mediterranean diet among GCAT participants we used the 14-item Mediterranean Diet Adherence Screener, a validated questionnaire that can be used in large epidemiological studies.[35] Additionally, a brief semiquantitative Food Frequency Questionnaire (FFQ), containing the food groups most eaten in Spain is used to assess total energy and macronutrient intake. A Spanish validated full-length FFQ (>128 questions), with a time frame referring to the previous 12 months, will be used during the first follow-up to assess dietary intake.[36]

Medical history, drug use and gender-specific information assessment

The medical history survey includes, among others, questions on current self-perceived health status (very good, good, fair, bad and very bad), minimum and maximal weight during the last 5 years and weight at birth. Current mental health is being assessed using the brief version of the Mental Health Inventory,[37] and asthma is being evaluated by using a categorical and continuous asthma score.[38] Specific questions for anti-inflammatory drugs and vitamin/mineral supplements use are also requested. Moreover, participants are asked whether a doctor has ever diagnosed 27 different diseases (table 1). If the answer is positive, then participants are asked for

Table 1 List of self-reported diseases at baseline for both men and women

ICD-9-CM* code	Conditions	Prevalent cases n (%)
272.0	Hypercholesterolaemia or Hypertriglyceridaemia	3456 (18.73)
995.3	Allergies	3132 (16.97)
401.9	Hypertension	2771 (12.31)
346.90	Migraine disorders	1614 (8.75)
472.0	Rhinitis	1415 (7.67)
311	Depression disorder	1193 (6.46)
493.90	Asthma	985 (5.34)
692.9	Eczema	872 (4.73)
41.86	*Helicobacter pylori* infection	855 (4.63)
569.0	Colon and/or rectal polyps	727 (3.94)
696.8	Psoriasis	688 (3.73)
250.00	Diabetes mellitus	613 (3.32)
733.00	Osteoporosis	578 (3.14)
714.9	Arthritis	552 (2.99)
199.1	Cancer	405 (2.19)
Z14	Inborn genetic diseases	172 (0.93)
496	Chronic obstructive pulmonary disease	89 (0.48)
558	Chronic colitis	53 (0.29)
434	Stroke	51 (0.28)
573.3	Chronic hepatitis	38 (0.21)
410.90	Myocardial infarction, heart attack	36 (0.19)
413	Coronary heart disease/angina pectoris	35 (0.19)
710	Lupus erythematosus	30 (0.16)
560.89	Crohn's disease	19 (0.10)
295.90	Schizophrenia	17 (0.09)
331	Alzheimer's disease/dementia	2 (0.01)
332	Parkinson's disease	1 (0.01)

*The International Statistical Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM).

the age at diagnosis and current drug use (drug name, frequency of use and age at first use). Drugs are being classified according to the Anatomical Therapeutic Chemical (ATC) Classification System.[39] The query also contains exhaustive questions on family history of diseases (parents, brother(s)/sister(s), and son(s)/daughter(s)).

The specific women's health questions are specially designed to assess a wide range of information on menstrual and reproductive history (including exogenous hormone use), and to study probable health variations in middle-aged women. Men are asked if they have ever been diagnosed with the most common prostate diseases (prostatitis or benign prostatic hyperplasia) and if they have ever taken any drug to treat them. Additionally, gender-specific screening programme participation is asked for both men and women.

Baseline environmental exposure assessment

Residences of all subjects will be geocoded, and a geographical information system based approach will be applied to evaluate environmental exposures (proximity to natural/green spaces, urban structure, air pollution, noise, temperature and artificial light at night) using existing information such as Urban Atlas, Corine Land Cover, Strategic Noise Maps (European Parliament directive 2002/49/CE and the Spanish Law 37/2003), Landsat images, International Space Station (ISS) images, among other available data.

Anthropometric measurements

Weight, height, waist and hip circumference (WC, HC; respectively), systolic and diastolic blood pressures, and heart rate are measured on all participants by trained personnel using the same protocol in all BST recruitment centres. Height is measured with a stable stadiometer (stable stadiometer for mobile height measurement, Seca 217, SECA UK), and weight is measured with electronic flat scales (bearing capacity of up to 200 kg, Seca 813, SECA, UK). Body circumference is measured using a measuring tape (ergonomic circumference measuring tape, Seca 201, SECA, UK) based on the WHO STEPwise Approach to Surveillance (STEPS) protocol.[40] WC is taken at the midpoint between the lower margin of the last palpable rib and the top of the iliac crest (expressed in cm). HC is taken at the maximum circumference over the buttocks (expressed in cm). All anthropometric measurements are performed in light clothing and twice. In case of discordances a third measure is taken. The average of the measures will be used for data analysis. A digital automatic blood pressure monitor (HEM-705CP, Omron Corporation, Tokyo, Japan) is used to measure systolic and diastolic blood pressures and heart ratio, which are displayed simultaneously. Participants are asked to sit and rest at least 5 min before taking three measurements (left arm). Interviewers wait 2–3 min between measurements. The average of the second and third measures will be used for data analysis as suggested by the WHO STEP protocol.[40] Body mass index (BMI; kg/m^2), WC and waist-to-hip ratio (WHR) will be categorised based on WHO guidelines.[41]

Biobanking of samples

For each participant, blood plasma, blood serum and white blood cells are collected at baseline according to standardised procedures (table 2). Different preservatives are used to collect blood specimens according to the downstream application.[42 43] The leucocyte residue (LR) is the main source for DNA extraction. The LR is a highly concentrated buffy coat obtained after the ordinary blood donation (blood bag of 480 mL). In those cases where blood donation is not possible or acceptable, an EDTA tube (10 mL) is obtained and used for DNA

Table 2 Suitability and processing of collected samples

Study purpose	Fraction sample	Vacutainer tube	Volume mL	Transport T°C	Time to PMPPC	Aliquots n (T°C)	Control assay*
Genomic/epigenomic	Buffy coat	EDTA	10	4	max 24 hours	2 (–80)	SNP array, qPCR, PCR, STR
	Highly concentrated buffy coat	Blood bag	480	18	max 48 hours	2 (–80)	SNP array, qPCR, PCR, STR
Proteomic/epigenomic	Plasma	PST	4.5	4	max 24 hours	4 (–80)	–
	Serum	SST	5	4	max 24 hours	4 (–80)	Circulating microRNAs integrity analysis
Functional/cell line	DMSO blood	ACD	6	18	–	2 (N_2)	EBV cell transformation and immortalisation

*Suitability downstream analysis performed in collected samples.
ACD, anticoagulant citrate dextrose; DMSO, dimethyl sulfoxide; EBV, Epstein-Barr virus; PMPPC, Program of Predictive and Personalized Medicine of Cancer; PST, plasma separation tube; qPCR, quantitative PCR; SNP, single nucleotide polymorphism; SST, serum separation tubes; STR, short tandem repeat.

extraction. Three vacutainer tubes with plasma separation tube (PST), serum separation tube (SST) and anticoagulant citrate dextrose (ACD) blood preservative are additionally collected for plasma, serum and viable cells, respectively.

The quality and quantity of the biospecimen samples is an important concern due to the GCAT long-term objectives. The suitability for maximising downstream applications is assured by developing an efficient recruitment infrastructure, by setting up a quality control tracking plan and by following dedicated procedures.

The GCAT Study has a centralised model, with two central laboratories (BST and PMPPC) and the recruitment satellites centres. Every day, samples are transported from all recruitment centres to the BST headquarters. From the ACD tubes, two aliquots of 1 mL are cryopreserved with dimethyl sulfoxide in cryogenic vials. These vials are stored in two independent liquid nitrogen tanks (–196°C) at the BST local repositories. The blood bag is also processed at the BST headquarters laboratory, where the LR is aliquoted into 15 mL Falcon tubes. EDTA, PSTs and SSTs are shipped to the PMPPC laboratory 24 hours after blood extraction at 4°C, while LR Falcon tubes are shipped <48 hours after blood extraction at 4°C.

Once samples arrive at PMPPC they are processed for storage. EDTA tubes are centrifuged at 2500 g for 10 min at 4°C, and the buffy coat is manually separated and aliquoted in tubes with 2D Data-Matrix codification (2D tubes) that fit in a 96-well plate Society for Biomolecular Screening (SBS) standard format. The buffy coat is manually aliquoted in two 2D tubes of 0.5 mL. SSTs and PSTs are immediately centrifuged after blood extraction at recruitment centres (2000 g for 10 min at room temperature, after 30 min of clotting time). Further, plasma derived from PSTs is centrifuged for a second time at PMPPC (2000 g for 10 min at 4°C). From PSTs and SSTs, four aliquots of 0.45 mL in four 2D tubes are made. From LR-Falcon tubes, two aliquots of 0.5 mL in two 2D tubes are derived. All samples aliquoted at PMPPC are processed using an automated liquid handling system (TECAN robot), and aliquots are stored in two independent ultra-freezers at –80°C with a CO_2 backup system.

This recruitment plan standardises the collection methods to minimise sample variability. The preanalytical variability derived from extraction methods to storage is registered using a standard preanalytical code (SPREC). The SPREC code offers an unbiased resource to evaluate any unexpected downstream finding. Further, all samples are registered daily in a laboratory integrated management system (Abbott Informatics-STARLIMS) that allows complete sample processing traceability. The central GCAT laboratory located at PMPPC currently contains around 270 000 2D tubes from blood aliquots from 18 659 participants.

The BST headquarters laboratory routinely performs a viral/bacterial antigen exposure determination, including hepatitis B virus, hepatitis C virus, HIV I/II, human T-lymphotropic virus I/II, syphilis (Lues) and Chagas. These results are personally communicated to all GCAT participants by letter under internal protocols. In case of a positive result, the participant is encouraged to visit a medical doctor.

Omic studies

In an early pilot study, omics techniques (ie, genome, metabolome, epigenome) will be used to determine molecular profiles from 6550 participants (6400 unrelated and 50 family trios) table 3. These 6400 unrelated participants were randomly selected from the GCAT cohort with a 1:1 gender proportion.

General metabolomic characterisation and specific lipoprotein profile of all 5000 blood plasma samples are currently being analysed using a combined untargeted approach of nuclear magnetic resonance spectroscopy and mass spectrometry at the Centre for Omic

Table 3 Summary of total omic data as of 2017

Study purpose	Number of participants	Fraction sample	Platform		Analysed
Metabolomic profile	5000	Plasma	NMR MS	–	150 metabolites
Genotype	5459	Buffy coat	Infinium Multi-Ethnic Global (MEGAEX2) array	HiScan confocal scanner (Illumina)	2×10^6 SNPs, InDels
Whole genome sequencing	808	Buffy coat	Illumina TruSeq PCR free/Illumina paired-end SBS	HiSeq 4000 sequencer (Illumina)	30× coverage
Subexome	200	Buffy coat	Agilent Sureselect/ Illumina paired-end SBS	MiSeq (Illumina)	Custom multigene panel 126 genes 400× coverage
Epigenome	150*	Whole blood	Methylation EPIC 850K array	HiScan confocal scanner (Illumina)	Differentially methylated analysis at single site and regional levels (genes, CpG island, promoters, enhancers)

EPIC, European Prospective Investigation into Cancer and Nutrition; InDels, insertions-deletions; MS, mass spectrometry; NMR, nuclear magnetic resonance; SBS, sequencing by synthesis; SNP, single nucleotide polymorphism.
*Current acquisition.

Sciences-Centre Tecnologic de Catalunya (COS-EU-RECAT) in Reus, Tarragona, Spain (online supplementary table1).

From the 6400 unrelated participants, 5459 genomic profiles have been characterised by comprehensive genotyping. Genome-wide genotypes have been generated using Illumina Infinium SNP-bead array technology. We chose the Multi-Ethnic Global (MEGAEX, V.2) consortium array, a multipurpose, multiethnic genotyping array with two million selected markers (including previously described germline mutations, insertions-deletions (InDels) and SNPs).[44] We have strictly followed the standard manufacturer recommended automated protocol for the Infinium HTS Assay scanned with a HiScan confocal scanner (Illumina, San Diego, California, USA). Genome Studio V.2011.1 has been used for raw data analysis. Genotyping was performed at the Genomics and Bioinformatics Unit of the PMPPC Institute for Health Science Research Germans Trias i Pujol, in Badalona, Spain.

A pilot family study including 50 related participants (parents and at least one offspring) is being conducted to reveal the role of DNA methylation as a key mechanism of heritability of chronic diseases. DNA methylation epigenomic profile of whole blood samples will be determined by Infinum Methylation EPIC 850K bead array assay.

Additionally, the entire genome of 808 participants will be sequenced with an overall coverage of 30×, using paired end sequencing by synthesis (SBS) on a HiSeq 4000 sequencer from Illumina (Illumina, San Diego, California, USA). Methylation analysis will be performed at the Genomic and Bioinformatics platform at PMPPC, and whole genome sequencing at the National Center for Genomic Analysis (CNAG-CRG) in Barcelona, Spain.

Two hundred participants will have overlapping array and sequencing characterisation, and will be further analysed for somatic genetic variance in hereditary cancer genes through high read depth targeted-subexome sequencing approach.

Active and passive follow-up

Participants will be followed for 20 years after recruitment. At the beginning of 2017, all GCAT participants received a newsletter by email acknowledging their participation and providing a brief explanation of the study status, the goals achieved and the future plans. The first active follow-up will start in 2018, and is planned to be biannual. Those participants who have been followed during at least 2 years received an electronic web-based epidemiological questionnaire (only accessible through a personalised link) to update or complement baseline information. Two reminders are planned to be sent in case of non-response (still ongoing). The follow-up survey was mainly designed to capture changes in health status, lifestyle (ie, smoking, physical activity, alcohol intake), dietary habits (validated full-length FFQ), circadian rhythm, shift-work and workplace environment (to study occupational diseases), among others.

Deceased participants during follow-up (end point ascertainment) will be identified by contrasting the data provided by the Spanish National Statistics Institute (www.ine.es). The National death statistics data are assembled following the WHO criteria, thus, all causes of death are classified according to the International Classification of Diseases (ICD; http://www.who.int/classifications/icd/en/).

The region of Catalonia has an advanced and highly developed healthcare system throughout the territory.

The GCAT Study has established a collaboration with the Catalan Health Department in order to have access to the EHRs of the Catalan Public Healthcare System. This registry comprises a huge amount of longitudinal clinical and personal information to promote EHR-driven research (ie, disease diagnosis, test reports, billing data, treatments, drug dosage/prescription, imaging data, biochemical analyses).[45 46] The EHR access protocol guarantees data confidentiality. Therefore, in an anonymous manner, the EHR information will be merged with the self-reported information that GCAT participants contributed at baseline. The EHR access will also allow us to follow participants during a long period of time, and to obtain a 5-year period retrospective health data.

Sample size and statistical power

At the end of 2017, the GCAT Study will have recruited 20 000 participants, and will be one of the largest prospective cohort studies in Spain. This will provide a powerful approach platform to study a wide range of complex diseases and related traits. In the early phases, incident cases will be included and analysed as part of a network of large cohort consortiums. Prevalent cases identified at baseline will provide an opportunity for early results based on several pathologies and related traits (table 1).

Statistical power for genetic associations is usually expressed by the number of estimated cases of the diseases of concern and the assumptions on the expected underlying genetic model. Based on the expected 20 000 participants, considering common conditions with 2.5%–5% prevalence at baseline, a case size of 500 individuals in a case-control study design (with 1:4 ratio), a power of 80%, an alpha level of 0.05, and under an additive genetic model (genetic power calculator), the minimum detectable statistically significant RR for low frequency variants (<5%) in complete linkage disequilibrium will be in the range of 1.5–2. Sample size increases by increasing the number of tested genetic markers to detect similar RR under same assumptions.[47] For other approaches, such as metabolite analyses, higher RRs are expected, being able to detect variation in metabolite concentration from 1% (r^2 >0.01) for a sample between 1000 and 5000 individuals, with >150 metabolites and 1×10^6 SNPs.[48] The size of the GCAT Study was initially settled considering the number of new cases expected to occur in the cohort along with the magnitude of the effect (RR) to be identified, as well the exposure prevalence;[21] however, this is relative and not unarguable when considering such a global approach.

Data management and analysis plan

Data management

The GCAT Study prospectively assembles data on lifestyle and dietary related risk factors. First, all epidemiological data collected at baseline will pass through a quality control process to ensure validity before examining the final data set. All data are collected with Onyx and are stored with Opal (the OBiBa's core database application for epidemiological studies),[49] and are housed in a secure high-performance computing and storage system at PMPPC.

Epidemiological and omic data analysis plan

As has been described before, the principal objective of GCAT is to prospectively investigate the association between epidemiological and genetic risk factors and different cancer sites and chronic diseases. Thus, several study designs such as cohort studies, nested case-control studies, cross-sectional studies, retrospective observational studies and studies based on routine data are planned (table 4). As a consequence, different statistical approaches will be used.

Molecular profiles will be linked to epidemiological data and personal EHRs to evaluate clinical associations (ie, cancer, cardiovascular, respiratory and neurological diseases, metabolic syndrome, and height). Outcomes of interest will evolve throughout the lifetime of the GCAT project.

GCAT genomic analysis will be used to characterise rare and low frequency variation in the Catalan-Spanish population. GCAT genomic profiles will be used to build genomic maps including both structural and sequence variations, and to create a population-specific sequence-based reference panel. Family data will be used for haplotype inference. A specific GCAT genome browser will provide interactive access to the project results.

Genomic quality control on raw genotyping will be performed with PLINK V.1.9 software. IMPUTE2 and SHAPEIT softwares will be used to impute untyped SNPs from sequence-based reference panels. Sequence data analyses include comprehensive quality control and the alignment reference genome (hg37) with GEM3.[50] GATK will be used to identify variants, annotate variants to gene (Ensembl) and analyse the in silico predicted functional impact (PhyloP, PolyPhen2, MutationTaster, CADD and GTEx), and population frequency (dbSNP, 1000GP, ExAC and Centro Nacional de Análisis Genómico-Centre for Genomic Regulation (CNAG-CRG) internal database).

Common and rare and structural genetic variant contribution will be analysed for heritable identified traits (biological or biomedical) with different predictive architectures. Genetic contribution to selected traits will be first analysed by GWAS for each variant using a multivariate logistic regression analysis. Variant effect size and p values will be derived. Whole genomic profile will be used for phenotype wide association analysis based on comprehensive clinical data from personal EHRs. The impact of population admixture will be analysed for clinical relevance based on population history.

Metabolomic and genomic integration will be performed to identify underlying genetic variants, as well as environmental factors that influence metabolites. Plasma metabolite profiles will be analysed for pathway analysis and diagnostic biomarker identification, and then metabolic quantitative trait analysis will be conducted to identify heritable endophenotypes for selected traits.

Table 4 Summary of all available Genomes for Life (GCAT) data

Data type	Number of participants	Details	Date of acquisition	Date available for research
Baseline assessment	Whole cohort	Questionnaire, physical measures, samples	2014–2017	2018
Repeat of baseline assessment	Whole cohort	Questionnaire follow-up every 2 years	2018	2019
Genotyping (baseline samples)	5459 (GCATcore)	Dense genotyping array with 666 695 markers after quality control (see figure 2)	2016	2018
Genotyping extended (baseline samples)	5459 (GCATcore)	Dense genotyping map with 15 078 461 variants (see figure 2) by in silico imputation (IMPUTE)	2017–2018	2018
Food frequency web questionnaire (follow-up)	Whole cohort	Participants are invited by email to provide additional information about diet; estimates of nutrient intake	2017–2018	2018
Biochemical assay (baseline samples)	6000	Glycated haemoglobin (haemoglobin A1c)	2016–2017	2018
Metabolome (baseline samples)	5000 (GCATcore)	Biomarkers with known disease association (lipids and vascular disease)	2017–2018	2018
Chronotype web questionnaire (follow-up)	Whole cohort	Participants are invited by email to provide additional information (ie, sleep behaviour, circadian rhythm, and work shift)	2017–2018	2018
Exposome (baseline)	Whole cohort	Map of environmental exposures acquired with geographical information system (GIS) technology	2017–2018	2018
Other web-based questionnaire data (follow-up)	Whole cohort	Participants are invited by email to provide additional information via web about working places. Information will be integrated with exposome assessment	2017–2018	2018
Exome	200 (GCATcore)	Clinic custom exome of hereditary cancer in 126 hereditary cancer genes (400×)	2017	2018
Whole-genome sequencing	808	30× whole genome sequencing from 1000 volunteers, 20% from GCATcore	2017–2018	2018
Epigenome	150	DNA methylation epigenomic profile using Infinum Methylation EPIC 850K beadarray assay	2018	2019
Health record linkage				
Primary care	Whole cohort	ICD/ATC/OPCS procedures/laboratory	2017–2018	2018
Death registrations	Whole cohort	ICD-coded cause specific mortality	2017–2018	2018
Hospital inpatient	Whole cohort	ICD/ATC/OPCS procedures/laboratory	2017–2018	2018
Hospital outpatient	Whole cohort	ICD (few)/OPCS	2018	2018
Other	Whole cohort	National mental healthcare/ national social healthcare	2018	2018

ATC, Anatomical Therapeutic Chemical Classification System; EPIC, European Prospective Investigation into Cancer and Nutrition; ICD, International Statistical Classification of Diseases; OPCS, Classification of Interventions and Procedures.

FINDINGS TO DATE

The GCAT Study is currently finishing the recruitment of participants (to be completed by December 2017). Among all GCAT participants, 59.2% are women and 83.3% of the cohort identified themselves as Caucasian/white. More than half of the participants have higher education levels, 72.2% are current workers and 42.1% are classified as overweight (BMI ≥25 and <30 kg/m^2)

Table 5 The Genomes for Life (GCAT) Study: summary of baseline characteristics

Characteristics	Values
Continuous variables	**Mean (SD)**
Age	51.03 (7.05)
Heart rate	74.47 (11.12)
Diastolic blood pressure	78.56 (9.71)
Systolic blood pressure	123.54 (15.28)
Age at menarche (among women)	12.38 (1.55)
Age at menopause (among women)	48.56 (4.74)
Age at voice change (among men)	14.7 (2.1)
Age at beard change (among men)	16.0 (2.6)
Categorical variables	**n (%)**
Gender	
Male	7471 (40.5)
Female	10 918 (59.2)
Missing	62 (0.3)
Marital status	
Married	10 703 (58.0)
Divorced/separated	2159 (11.7)
Domestic partner	1142 (6.2)
Single	1887 (10.2)
Widow/widower	521 (2.8)
Missing	2039 (11.1)
Education level	
Without studies	73 (0.4)
Elementary education	2104 (11.4)
Secondary education	4519 (24.5)
Professional higher education	2037 (11.0)
Secondary postdegree professional programme	2594 (14.1)
College	6772 (36.7)
Missing	352 (1.9)
Ethnicity	
White, Caucasian	15 363 (83.3)
Hispanic, Latin	2803 (15.2)
Black	14 (0.1)
Maghrebin	14 (0.1)
Gipsy	10 (0.1)
Asian	1 (0.0)
Other	18 (0.1)
Missing	230 (1.2)
Working status	
Employed	13 327 (72.2)
Not working/employed	1796 (9.7)
Retired	1255 (6.8)
Home maker	1110 (6.0)
Student	52 (0.3)
Laboral impairment	376 (2.0)
Volunteer or unpaid work	126 (0.7)
Other	206 (1.1)
Missing	203 (1.1)
Smoking status	
current, <=15 cig/day	2469 (13.4)
current, 26+cig/day	148 (0.8)
current, unknown	318 (1.7)
current, 16–25 cig/day	752 (4.1)
former, quit<=10 years	2196 (11.9)
former, unknown	153 (0.8)
former, quit 11–20 years	2392 (13.0)
former, quit 20+ years	1973 (10.7)
missing	853 (4.6)
never	7197 (39.0)
Alcohol consumption	
never or less than once a month	4402 (23.9)
once per month	1048 (5.7)
from 2 to 3 times per month	2202 (11.9)
once per week	3061 (16.6)
from 2 to 3 times per week	3454 (18.7)
from 4 to 6 times per week	1059 (5.7)
once per day	1963 (10.6)
two or more times per day	1036 (5.6)
missing	226 (1.2)
Mediterranean Diet Adherence (PrediMed Score)	
Low	2159 (11.7)
Medium	12 904 (70)
High	2893 (15.7)
Missing	495 (2.7)
Health status	
Very good	3124 (16.9)
Good	13 080 (70.9)
Regular	1960 (10.6)
Bad	126 (0.7)
Very bad	20 (0.1)
Missing	141 (0.8)
Adopted	
Yes	60 (0.3)
No	18 243 (98.9)
Missing	148 (0.8)
Body mass index	
Underweight	47 (0.2)

Table 5 Continued

Characteristics	Values
Normal weight	6083 (33)
Overweight	7761 (42.1)
Obese	4562 (24.7)
Missing	89 (0.5)
Women related health	
Oral contraceptive use	
Never	2351 (21.5)
Ever	8404 (77)
Missing	163 (1.5)
Hormone replacement therapy (HRT) use	
Never	9280 (85)
Ever	1317 (12.1)
Missing	321 (2.9)
Men related health	
Prostate diseases	660 (8.8)

Two types of variables, continuous (presented in mean (SD)) and categorical (which are presented in n(%)) are shown in bold.

(table 5). The first active follow-up of the first volunteers entering the study will begin in January 2018 and will end in March 2018.

Genomic characterisation using array-based technology of subcohort (GCATcore data release August 2017), (figure 2).

The results of the study will be published in international peer-reviewed journals and presented at national and international congresses and conferences. Preliminary data have already been analysed and presented.[51 52]

DISCUSSION

One of the major strengths of the GCAT Study is its prospective design, and that it is one of the largest EHR linked-cohort studies in Spain with a deep genome-wide characterisation. In addition, blood plasma, blood serum and white blood cells were collected and stored at baseline for each participant. All epidemiological and anthropometric data have been annotated using international codification to allow data exchange between national and international studies, and to be part of a network of large cohort consortiums, as a global strategy on health (ie, Genomes of England). Further, detailed information on clinical and health status is available for each participant, as the GCAT Study has access to EHRs of the Catalan Public Healthcare System. The GCAT Study offers a unique opportunity to integrate epidemiological, environmental, EHR and omic factors to investigate the aetiology of chronic diseases. Molecular pathological epidemiology is a new research area that integrates different fields with the aim to study phenotypes of any disease using molecular pathological analyses.[53 54] Analyses with germline genomic, epigenomic and metabolomic data will provide new results and derive new scientific knowledge for public health interventions (primary and secondary prevention) and towards precision medicine for personalised prevention.

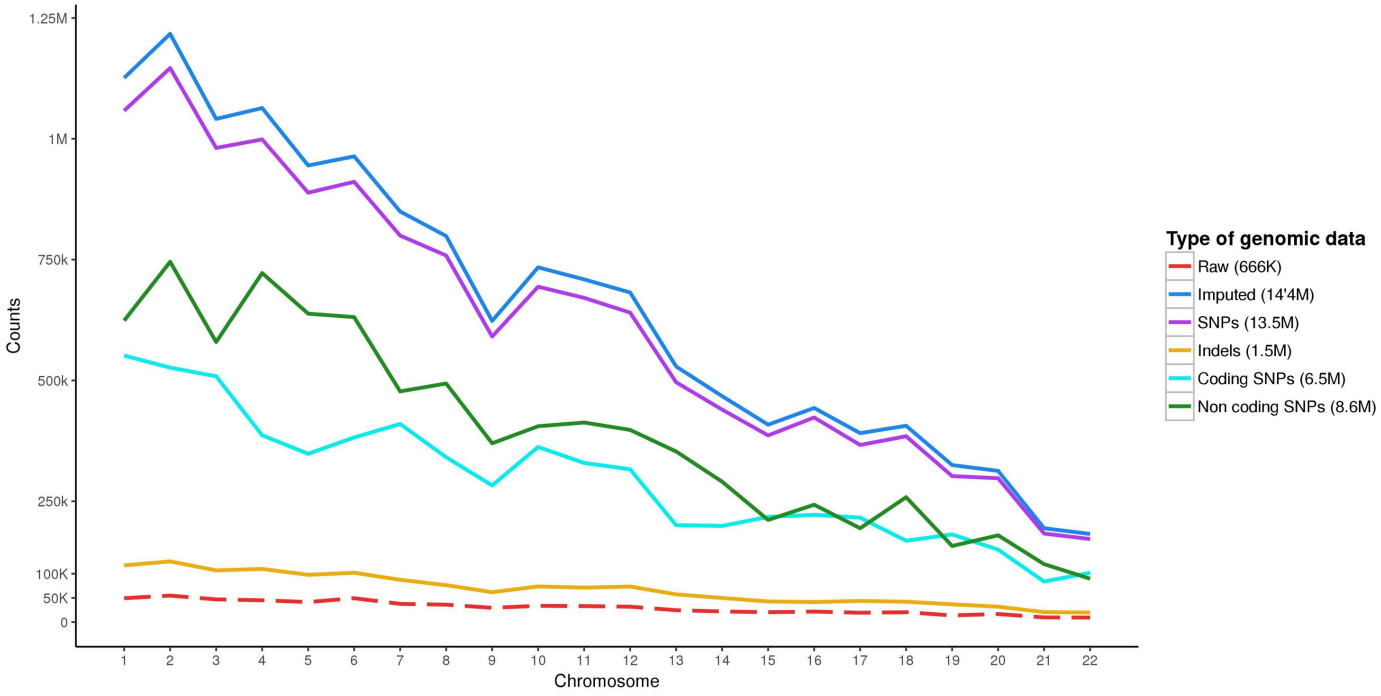

Figure 2 Number of variants included in the GCATcore by chromosome and type of genomic data (first GCATcore release August 2017). The legend shows between parentheses the total number of variants for raw data (genotyping before in silico imputation), imputed data, SNPs, Indels and SNPs in coding and non-coding regions. GCAT, Genomes for Life; InDels, insertions-deletions; single nucleotide polymorphisms.

There are a number of challenges that should be acknowledged. The GCAT Study was not designed to achieve a representative sample of the Catalan population, since non-representativeness does not usually interfere with scientific inference.[55 56] The GCAT recruitment process may have introduced selection bias in our study; however, the recruitment was performed through the BST agency to enhance participation, and assuring a long-term follow-up (20 years). As stated before, our cohort participants are mainly health conscious; nevertheless, these participants are more likely to participate in intervention studies (which are planned at later stages) to evaluate, for instance, behavioural and lifestyle habit changes. Further, with a global disease approach, the sample size will be a limitation to test for genetic associations in any condition even in larger cohorts; nonetheless, the deep phenome characterisation of the GCAT cohort will allow the implementation of systems biology approaches. As the analyses expand (including copy number variants, rare alleles and other types of methods) more associations will be identified, leading to an increase in knowledge of the influence of genomic structure and function on health and common diseases.

COLLABORATION

One of the GCAT characteristics is that it has an open protocol that will enable future study designs or procedures (ie, family studies, intervention studies). Epidemiological data and biological samples are available for external researchers. Genotypes (SNPs and InDels) and sequence variation data will be sent in a multisample variant call file to facilitate further analyses. Qualified researchers who fulfil ethical and scientific requirements can submit an application form with their personal information, a brief summary of the project and specific data/material requested. The GCAT scientific committee will evaluate the proposals. Before sending biological material and/or data, a data/material transfer agreement form will be signed among partners to ensure right and duties. All information regarding the ethical legal social issues, questionnaire contents and available data can be found at www.genomesforlife.com.

CONCLUSIONS

The GCAT Study is a long-term genomic, environmental and lifestyle cohort project that aims to evaluate and track multiple pathologies as well as biologically related traits. Therefore, the GCAT Study offers a unique opportunity to integrate diverse data to allow the identification of novel relations among different biomarkers and conditions. Results may lead to the development of new genetic, genomic, epigenomic and proteomic diagnoses and screening tests, as well as new public health recommendations.

Author affiliations
[1]Genomes for Life -GCAT lab Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona, Spain
[2]Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO-IDIBELL), Hospitalet del Llobregat, Spain
[3]Banc de Sang i Teixits (BST), Barcelona, Spain
[4]Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona, Spain
[5]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO-IDIBELL), Hospitalet del Llobregat, Spain
[6]CIBER Epidemiología y Salud Pública (CIBERESP), Hospitalet del Llobregat, Madrid, Spain
[7]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain

Acknowledgements The authors thank all the GCAT participants and all BST members for generously helping with this research. The authors also thank the PMMPPC-IGTP personnel David Piñeyro, Laia Ramos, Raquel Pluvinet and Susanna Aussó from the Genomics and Bioinformatics Units for genotyping support, Ivo Gut and CNAG-CRG personnel for sequencing support, Núria Canela and COS-EURECAT for metabolomic analysis support, Harvey Evans, communications manager, and Victor Bonet and Hardeep Kaur for data entry. The authors also thank Marta Guindo and David Torrents for their help on the genotype imputation and the use of the MareNostrrum in the Barcelona Supercomputing Center (BSC), Isabel Fortier and Vincent Ferretti for their helpful insights on the eGCAT design, the Maelstrom Research Group (Research Institute of the McGill University Health Center Montreal, Canada) for their support on the eGCAT customisation, and Manolis Kogevinas (CREAL-IsGlobal, Barcelona) for contribution to exposoma design.

Contributors All authors contributed to feedback of the manuscript. All authors played an important role in implementing the study protocol.Conception and design: RdeC, VM, MP, EJD, MO-S. Development of methodology: RdeC, VM, EJD, AC, MO-S, XD, IG-F, LS, JV, LP. Writing, review and/or revision of the manuscript: MO-S, MV, TA, RdeC, VM, EJD, MP, AC, LS, XD, IG-F. Administrative, technical or material support: MP, RdeC, VM, EJD, AC, MO-S, XD, IG-F, JV, LP. Study supervision: RdeC, VM, MP, EJD, MO-S.

Funding This work was supported by Acción de Dinamización del ISCIII-MINECO (ADE 10/00026), by the Ministry of Health of the Generalitat of Catalunya and by Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (SGR 1269 and 1589) and by the Catalan Government DURSI (grant 2014SGR647). Dr Rafael de Cid is the recipient of a 'Ramón y Cajal' (RYC) action (RYC-2011-07822) from the Spanish Ministry of Economy and Competitiveness. The Project is coordinated by the Germans Trias i Pujol Research Institute (IGTP), in collaboration with the Catalan Institute of Oncology (ICO), and in partnership with the central Blood and Tissue Bank of Catalonia (BST). IGTP is part of the CERCA Programme/Generalitat de Catalunya.

Competing interests None declared.

Patient consent Obtained.

Ethics approval The GCAT study was approved by the local Ethics Committee (Germans Trias University Hospital) in 2013.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement One of the GCAT characteristics is that it has an open protocol that will enable future study designs or procedures (i.e., family studies). Epidemiological data and biological samples are available for external researchers. Qualified researchers who fulfill ethical and scientific requirements can submit an application form with their personal information, a brief summary of the project, and specific data/material requested (www.genomesforlife.com/investigadors/daccess-documents/). The GCAT scientific committee will evaluate the proposals. Before sending biological material and/or data, a data/material transfer agreement form will be signed among partners to ensure rights and duties. All information regarding the Ethical Legal Social issues, questionnaire contents and available data can be found at www.genomesforlife.com/investigadors/. The results of the study will be published in international peer-reviewed journals and presented at national and international congresses and conferences. Summary data available could be consulted at http://www.genomesforlife.com/investigadors/en_gcat-summary-aggregate-data/.

REFERENCES

1. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015;386:743–800.
2. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 2016;388:1545–602.
3. World Health Organization. *Noncommunicable diseases progress monitor 2015*: World Health Organization Press, 2015.
4. World Health Organization. *Projections of mortality and causes of death, 2015 and 2030*: World Health Organization Press, 2014.
5. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, *et al*. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 2013;49:1374–403.
6. Galceran J, Ameijide A, Carulla M, *et al*. Cancer incidence in Spain, 2015. *Clin Transl Oncol* 2017;19:799–825.
7. Bloom D, Cafiero E, Jané-Llopis E, *et al*. *The global economic burden of noncommunicable diseases*: World Economic Forum, 2012.
8. Rappaport SM. Genetic Factors Are Not the Major Causes of Chronic Diseases. *PLoS One* 2016;11:e0154387.
9. Ottman R. Gene-environment interaction: definitions and study designs. *Prev Med* 1996;25:764–70.
10. Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. *Hum Hered* 2007;63:63–6.
11. MacArthur J, Bowler E, Cerezo M, *et al*. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;45:D896–D901.
12. Visscher PM, Brown MA, McCarthy MI, *et al*. Five years of GWAS discovery. *Am J Hum Genet* 2012;90:7–24.
13. Chang CQ, Yesupriya A, Rowell JL, *et al*. A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur J Hum Genet* 2014;22:402–8.
14. Ioannidis JP, Castaldi P, Evangelou E. A compendium of genome-wide associations for cancer: critical synopsis and reappraisal. *J Natl Cancer Inst* 2010;102:846–58.
15. Gorlov IP, Gorlova OY, Sunyaev SR, *et al*. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 2008;82:100–12.
16. van Dijk EL, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014;30:418–26.
17. Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health* 1995;16:61–81.
18. Thomas F. *Handbook of Migration and Health*: Edward Elgar Publishing, 2016.
19. Staszewski J. Migrant studies in alimentary tract cancer. *Recent Results Cancer Res* 1972;39:85–97.
20. Organization WH. *Global health risks: mortality and burden of disease attributable to selected major risks*: World Health Organization, 2009.
21. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 2008.
22. Idescat I. Statistical Yearbook of Catalonia. Population. Provinces. http://www.idescat.cat/pub/?id=aec&n=245&lang=en (accessed 11 Jan 2017).
23. Bälter O, Bälter KA. Demands on web survey tools for epidemiological research. *Eur J Epidemiol* 2005;20:137–9.
24. Doiron D, Burton P, Marcon Y, *et al*. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol* 2013;10:12.
25. Fortier I, Burton PR, Robson PJ, *et al*. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;39:1383–93.
26. Instituto Nacional de Estadística. Clasificación Nacional de Ocupaciones. CNO-11. http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177033&menu=ultiDatos&idp=1254735976614 (accessed 3 Mar 2017).
27. ISCO - International Standard Classification of Occupations. http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm (accessed 3 Mar 2017).
28. Heatherton TF, Kozlowski LT, Frecker RC, *et al*. Measuring the heaviness of smoking: using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. *Br J Addict* 1989;84:791–800.
29. Rodríguez-Martos Dauer A, Gual Solé A, Llopis Llácer JJ. [The "standard drink unit" as a simplified record of alcoholic drink consumption and its measurement in Spain]. *Med Clin* 1999;112:446–50.
30. Rehm J, Room R, Monteiro M, *et al*. *Alcohol use: Chapter 12*: Comparative Quantification of Health Risks WHO, 2012:0959–1108.
31. Mäkelä P, Gmel G, Grittner U, *et al*. Drinking patterns and their gender differences in Europe. *Alcohol Alcohol Suppl* 2006;41:i8–i18.
32. Peters T, Brage S, Westgate K, *et al*. Validity of a short questionnaire to assess physical activity in 10 European countries. *Eur J Epidemiol* 2012;27:15–25.
33. Ainsworth BE, Haskell WL, Whitt MC, *et al*. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc* 2000;32:S498–S516.
34. Cust AE, Smith BJ, Chau J, *et al*. Validity and repeatability of the EPIC physical activity questionnaire: a validation study using accelerometers as an objective measure. *Int J Behav Nutr Phys Act* 2008;5:33.
35. Schröder H, Fitó M, Estruch R, *et al*. A short screener is valid for assessing Mediterranean diet adherence among older Spanish men and women. *J Nutr* 2011;141:1140–5.
36. Fernández-Ballart JD, Piñol JL, Zazpe I, *et al*. Relative validity of a semi-quantitative food-frequency questionnaire in an elderly Mediterranean population of Spain. *Br J Nutr* 2010;103:1808–16.
37. Rumpf HJ, Meyer C, Hapke U, *et al*. Screening for mental health: validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. *Psychiatry Res* 2001;105:243–53.
38. Pekkanen J, Sunyer J, Anto JM, *et al*. Operational definitions of asthma in studies on its aetiology. *Eur Respir J* 2005;26:28–35.
39. Organization WH. *The anatomical therapeutic chemical classification system with defined daily doses (ATC/DDD)*. Norway: WHO, 2006.
40. World Health Organization. *WHO STEPwise approach to surveillance (STEPS)*, 2008.
41. *Obesity: preventing and managing the global epidemic. Report of a WHO consultation*: World Health Organ TechRepSer, 2000;894: i–253.
42. Elliott P, Peakman TC. UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 2008;37:234–44.
43. Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. *Int J Epidemiol* 2008;37 Suppl 1(Suppl 1):i2–i6.
44. Bien SA, Wojcik GL, Zubair N, *et al*. Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS One* 2016;11:e0167758.
45. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13:395–405.
46. Marimon-Suñol S, Rovira-Barberà M, Acedo-Anta M, *et al*. [Shared electronic health record in Catalonia, Spain]. *Med Clin* 2010;134 Suppl 1(Suppl 1):45–8.
47. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform* 2012;10:117–22.
48. Nicholson G, Rantalainen M, Li JV, *et al*. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet* 2011;7:e1002270.
49. OBiBa: Open Source Software for BioBanks. http://www.obiba.org/ (accessed 13 Mar 2017).
50. Marco-Sola S, Sammeth M, Guigó R, *et al*. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 2012;9:1185–8.
51. Obón-Santacana M, Vilardell M, Carreras A, *et al*. GCAT|Genomes for Life: A prospective cohort study of the genomes of Catalonia [abstract]. In. *European Human Genetics Conference 2016*. 2016. Barcelona. Spain: ESHG 2016, 2016. Abstract nr P18.042.
52. Galván-Femenía I, Graffelman J, de Cid R, *et al*. Graphical tools for estimating family relationships [abstract]. In. *European Human Genetics Conference 2016*. Barcelona. Spain: ESHG 2016, 2016. Abstract nr P18.061.
53. Hamada T, Keum N, Nishihara R, *et al*. Molecular pathological epidemiology: new developing frontiers of big data science to study etiologies and pathogenesis. *J Gastroenterol* 2017;52:265–75.
54. Ogino S, Nishihara R, VanderWeele TJ, *et al*. Review Article: The Role of Molecular Pathological Epidemiology in the Study of Neoplastic and Non-neoplastic Diseases in the Era of Precision Medicine. *Epidemiology* 2016;27:602–11.
55. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;42:1012–4.
56. Richiardi L, Pizzi C, Pearce N. Commentary: Representativeness is usually not necessary and often should be avoided. *Int J Epidemiol* 2013;42:1018–22.