2013-09-25

LifeGene : a large prospective population-based study of global relevance

Almqvist, Catarina; Adami, Hans-Olov; Franks, Paul W; Groop, Leif; Ingelsson, Erik; Kere, Juha; Lissner, Lauren; Litton, Jan-Eric; Maeurer, Markus; Michaëlsson, Karl; Palmgren, Juni; Pershagen, Göran; Ploner, Alexander; Sullivan, Patrick F; Tybring, Gunnel; Pedersen, Nancy L

Eur J Epidemiol. 2011 Jan;26(1):67-77.
http://doi.org/10.1007/s10654-010-9521-x
http://hdl.handle.net/10616/41729

LifeGene – a large prospective population-based study of global relevance

Catarina Almqvist MD PhD[1,2]
Hans-Olov Adami MD PhD[1,3]
Paul W Franks PhD[4,5]
Leif Groop MD PhD[5]
Erik Ingelsson MD PhD[1]
Juha Kere PhD[6]
Lauren Lissner PhD[7]
Jan-Eric Litton PhD[1]
Markus Maeurer MD PhD[8]
Karl Michaëlsson MD PhD[9]
Juni Palmgren PhD[1,10]
Göran Pershagen MD PhD[11]
Alexander Ploner PhD[1]
Patrick F Sullivan MD PhD[12]
Gunnel Tybring PhD[1]
Nancy L Pedersen PhD[1]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm Sweden
[2] Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm Sweden
[3] Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts USA
[4] Department of Public Health & Clinical Medicine, Section for Medicine, Umeå University Hospital Sweden
[5] Department of Clinical Sciences, Diabetes and Endocrinology Unit, Lund University, Lund Sweden
[6] Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge Sweden
[7] Department of Public Health and Community Medicine, University of Gothenburg, Sweden
[8] Department of Microbiology, Tumor and Cell Biology, KI and the Swedish Institute for Infectious Disease Control, Stockholm Sweden
[9] Department of Surgical Sciences, Uppsala University Sweden
[10] Department of Mathematical Statistics, Stockholm University Sweden
[11] Institute of Environmental Medicine, KI, Stockholm Sweden
[12] Department of Genetics, University of North Carolina at Chapel Hill USA

LifeGene – population-based cohort

Corresponding authors

Catarina Almqvist and Nancy L. Pedersen
Department of Medical Epidemiology and Biostatistics,
Box 281, Karolinska Institutet
SE-171 77 Stockholm SWEDEN
T +46 (0)8 524 87418
F +46 (0)8 31 49 75
E catarina.almqvist@ki.se and nancy.pedersen@ki.se

Abstract

Studying gene-environment interactions requires that the amount and quality of the lifestyle data is comparable to what is available for the corresponding genomic data. Sweden has several crucial prerequisites for comprehensive longitudinal biomedical research, such as the personal identity number, the universally available national health care system, continuously updated population and health registries and a scientifically motivated population. LifeGene builds on these strengths to bridge the gap between basic research and clinical applications with particular attention to populations, through a unique design in a research-friendly setting.

LifeGene is designed both as a prospective cohort study and an infrastructure with repeated contacts of study participants approximately every five years. Index persons aged 18-45 years old will be recruited and invited to include their household members (partner and any children). A comprehensive questionnaire addressing cutting-edge research questions will be administered through the web with short follow-ups annually. Biosamples and physical measurements will also be collected at baseline, and re-administered every five years thereafter. Event-based sampling will be a key feature of LifeGene. The household-based design will give the opportunity to involve young couples prior to and during pregnancy, allowing for the first study of children born into cohort with complete pre-and perinatal data from both the mother and father. Questions and sampling schemes will be tailored to the participants' age and life events. The target of LifeGene is to enrol 500,000 Swedes and follow them longitudinally for at least 20 years.

Key words [6 words]
Biobank, cohort study, epidemiology, prospective study, questionnaires, population genetics

Abbreviations

DLW	doubly labelled water
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
GWAS	genome wide association studies
ILI	influenza-like illness
PCR	polymerase chain reaction
WHO	World Health Organisation

Introduction

Background

In only a few years the methods used to study genetics of complex traits have evolved almost beyond recognition. The result has been unprecedented advances in our understanding of the genetic architecture of almost all common complex diseases. The main driving force behind these discoveries has been the rapid development and application of genome-wide association studies (GWAS). Although there has been remarkable recent progress in defining the genetic basis of common complex diseases, a new generation of epidemiological studies that undertake standardised, fine-scale phenotyping and exposure assessments of many thousands of individuals measured repeatedly during many years of follow-up will be required to maintain this trajectory. New approaches for storing, processing, and charactering the tissue samples collected in these individuals will be necessary, as will the application of novel, comprehensive, and multi-factorial statistical methods to analyse data harvested from these biosamples.

Worldwide, there are several large-scale genetic epidemiological studies such as the UK Biobank (1), deCODE in Iceland (2), Millennium in Japan (3), the Kadoorie study of Chronic Diseases in China (4) and CONOR/HUNT in Norway (5). These studies focus on chronic diseases diagnosed later in life, and some also collect detailed data on lifestyle factors. There are, however, very few large prospective studies focusing on disorders that emerge early in life, in some instances during infancy or early adolescence. LifeLines in Holland (6), a three-generation population-based study with a household recruitment approach differs from most other ongoing cohorts because data on diseases with onset early in life is collected. The Norwegian Mother and Child Cohort study (MoBa) (7), ALSPAC (the Avon Longitudinal Study of Parents and Children) (8) and the US National Child Study (9) have followed pregnant mothers from early pregnancy and their offspring throughout childhood, whilst other prospective birth cohort studies were brought together under the Global Asthma and Allergen European Network (Ga2len) (10). The Southampton Women's Survey (11) is of particular interest because it collected parental exposure data *before* the pregnancy and thus could assess associations

to perinatal and infant outcomes. Pre-pregnancy exposure data are also likely to be valuable in long-term studies of chronic disease later in life, although to date no such studies have been undertaken.

Sweden and other Nordic countries have several crucial prerequisites for comprehensive longitudinal biomedical research, due to the personal identity number, the universally available national health care system, continuously updated population and health registries and a scientifically motivated population. As a complement to the register-based epidemiologic tradition, Sweden is also at the forefront of developing and implementing information technology, cutting edge biotechnology and biobanking. In order to adequately address gene-environment interactions, acquisition of phenotype and exposure information repeatedly, biosamples and physiological assessment are important. A prospective study combining information on exposures as early as pre-conception and in early life throughout young and mid-adulthood, with the possibility to study a variety of diseases with onset relatively early in life and co-morbidities later in life will open up new research opportunities.

LifeGene is designed as a prospective cohort study with an infrastructure that allows repeated contacts of study participants approximately every five years, and short follow-ups annually. Recruitment of index people aged 18-45 years old who are invited to include their household members (other adults and any children) increases the opportunity to involve young couples prior to and during pregnancy, allowing for the first study of children born into cohort with complete pre- and perinatal data from both the mother and father. Other types of event-based sampling, i.e. data collection initiated as a result of a relevant event, such as an accident or influenza, is a key feature of LifeGene. A comprehensive web-based questionnaire comprised of multifaceted questions concerning phenotypes and exposures is administered at baseline, briefly followed-up annually, and re-administered every five years thereafter. Biosamples are collected and stored in a manner that facilitates the full range of ‘omics´, and physical measurements are also sampled repeatedly at five

year intervals. Questions and sampling schemes will be tailored to the participants' age and life events. The target of LifeGene is to enrol 500,000 Swedes and follow them for at least 20 years.

Methods and planning process

In the early phases of planning Life Gene, multiple workshops and plenary sessions with invited international and national experts in a variety of diseases, genetic epidemiology, environmental epidemiology, and molecular genetics were organised during which the optimal LifeGene sample design, research questions that could be addressed with a prospective cohort, biological measures and phenotype and exposures were discussed.

Separate working groups focusing on six broad phenotypic domains: infections, inflammation and allergy, malignant diseases, metabolic and cardiovascular disorders, musculoskeletal disorders and neuro-psychiatric disorders were established. In addition, a working group was established to discuss lifestyle factors such as physical activity and diet. Each working group was invited to present their area of research and the rationale for participation in LifeGene, along with preferred study design and optimal measures to include in questionnaires, biosamples and physical assessments. Three working groups focusing on infrastructure issues were also created: Biostatistics, biosampling and biobanking, and IT-infrastructure. Based on the combined recommendations of all groups, study design, recruitment and assessment schemes and protocols were established.

Infectious diseases

Outcomes of exposures to infectious agents are shaped by genetic background, exposure history to pathogens (or antigens in general, including vaccines), along with innate and adaptive immune response. Exposure to infectious agents has a broader impact on health and disease as compared to infections *strictu sensu*. Chronic inflammation, in part induced by infection, may impact on a number of diseases; prime examples of infections associated with malignant transformation are infections with Helicobacter pylori or human papilloma virus type 16. Perhaps the most recent report which links a retroviral infection with chronic fatigue syndrome shows that gauging infections entails far

more than just the examination of acute clinical infections. It also showed that retrospective analysis of old diseases may lead to the identification of novel pathogens (12). Novel and more sensitive methods to detect infectious pathogens as well as gauging relevant immune responses will help elucidate associations between infections and clinical syndromes.

Pathogenicity of infectious agents is in part determined by the intricate host-pathogen relationship, which can be captured in a LifeGene grid of household interviews and biological sampling. An event-based sampling study design allows linkage of clinically and biologically relevant events with the detection of infectious agents. An exercise in this context was proposed for the LifeGene pilot, which addressed Influenza-like-illnesses (ILI), described further on page 24.

Other information relevant to infectious and viral diseases, such as travel, sexual behaviour, and vaccination will also be collected.

Inflammation and allergy

Inflammation is a key factor in allergy and asthma as well as in rheumatic diseases, multiple sclerosis and other autoimmune diseases. It also plays an important role in cardiovascular, musculoskeletal and periodontal diseases. Important research questions include the role of factors pre-conception for disease development in the offspring, which may be mediated by epigenetic mechanisms. A LifeGene birth cohort design would also be uniquely suited for addressing questions related to fetal growth. Furthermore, research questions on gene-environment interactions are paramount and a wide variety of early life exposure and conditions are of interest, such as nutrition, infections, maternal stress, smoking and obesity, as well as exposure to allergens and ambient air pollution.

Special focus for protocol development was given to recruitment of pregnant mothers and their partners, including questionnaires and biosamples before and during pregnancy, as well as on extensive long-term follow-up of children born into the cohort. This group is of prime interest also for many other research questions, such as testing and providing mechanistic explanations for the

Barker hypothesis (13). Women in the LifeGene cohort who become pregnant were proposed to answer extensive questionnaires at gestational week 10-12 and at week 26-28, and biological samples to be taken using the same methodology as for the full cohort. Cord blood and maternal blood at delivery were also proposed and the child to be followed up with questionnaires at 4 months and then at 1 year of age, followed by yearly questionnaires. In subgroups additional biosamples may be obtained on several occasions during the first two years of life.

Malignant diseases

Progress in cancer epidemiology has been remarkable during the last decades. Several malignancies – including those predominantly caused by tobacco, alcohol and certain oncogenic infections (e.g. hepatitis virus B and C, human papilloma virus and *helicobacter pylori*) – are now largely preventable. At the same time, progress has been limited or non-existent in other areas. Indeed, the causes of major cancer sites and types remain enigmatic. Prostate cancer, lymphomas, leukaemia and testicular cancer are salient examples.

The cancer working group agreed that LifeGene would provide unparallel novel opportunities to reveal genetic and environmental causes of cancer. No other existing study combines the size of LifeGene with opportunities for repeated exposure measurement starting early in life, an extensive biobank allowing studies of biomarkers and genes with the convenience in Sweden of retrieving tumor tissue from pathology departments. As a corollary, LifeGene provides the opportunities to take a life course approach to cancer etiology; to accommodate emerging hypotheses through repeated exposure assessment by creative application of e-epidemiology; and to apply novel molecular techniques for refined – and perhaps etiologically more relevant – phenotypic subgrouping of cancer sites and types. The intense methodologic development in studies of gene-environment interactions will only make LifeGene more informative over time.

Metabolic and cardiovascular disorders

The current global increase of obesity and type 2 diabetes incidence has taken epidemic proportions, and even if this is almost certainly due to a global shift towards more obesogenic behaviours, such as physical inactivity and caloric dense diet, susceptibility to these conditions varies greatly from one person to the next and tends to segregate within families, irrespective of lifestyle. Furthermore, the response to lifestyle interventions differs markedly from one person to the next, which also reinforces the hypothesis that metabolic and cardiovascular disorders emerge through a complex interaction of genetic and environmental factors. LifeGene provides a unique opportunity to study such interactions in detail with the possibility to address primordial, early-life factors and the role of longitudinal changes for the development of these diseases.

The collection and storage of DNA, serum, plasma and urine allows for a core set of selected biomarkers, of which most are highly relevant for metabolic and cardiovascular disorders, to be analysed in front-end chemistry, and for subsequent analyses of genetic markers, proteins and metabolites in frozen specimens. As HbA1c is introduced for diagnosis of diabetes in many countries, measurement of HbA1c could provide a valuable estimate of the proportion of people with clearly abnormal glucose tolerance in the age group.

Musculoskeletal disorders

Musculoskeletal problems are commonly related to injuries, degenerative changes or from still unknown causes. With impaired functioning and pain as the dominant consequences for the individual, they bring high costs for the society (14).

The high incidence of osteoporotic fractures in Sweden, and other Scandinavian countries, is an observation that cannot readily be explained by known lifestyle or genetic determinants, climate or longevity. Osteoarthritis, back and shoulder pain are common reasons for disabling pain and sick leave. There is also emerging evidence for an interaction between bone and the endocrine system. LifeGene will provide opportunities to combine metabolic information with information on bone

density and fracture risk. LifeGene will examine changes in bone mineral density, bone structure and body composition in a subgroup of individuals. The diagnoses of clinically manifest arthritis can be obtained from Swedish registries, but LifeGene will provide an opportunity to detect pre-clinical biomarkers for the diseases, and to furthermore evaluate the relative importance of genetic susceptibility of arthritis compared to life-style behaviours.

Worldwide, injuries are one of the major causes of death, disability and health care consumption at all ages below 60. Sports, play, traffic accidents and intentional injuries (violence and self-harm) are the most common causes. A special interest will be to follow children from pre-puberty until they reach 25 years of age to evaluate genetic and environmental predictors for peak bone mass. Detailed questionnaire information on causes (event-based) and consequences of the injuries, along with genetic and lifestyle characterisation of individuals at high risk of severe or repeated injuries will be enabled with detailed prospective characterisation of participants at an especially high risk for the development of these diseases.

Neuro-psychiatric disorders and hearing

The etiology of most neuro-psychiatric disorders remains cryptic although extant data suggest that most have roots in childhood. The national health registers capture inpatient admissions relatively well although these constitute a small fraction of population burden. Patient charts from general medical settings seldom include psychiatric diagnoses. In the LifeGene neuro-psychiatric working group, initial focus was thus on which neuro-psychiatric disorders to consider and, second, which might be reasonable to assess with a web-based instrument. Using Swedish data on disability-adjusted life year estimates obtained from the World Health Organisation (WHO) Global Burden of Disease (15) as a guide, the working group defined a list of disorders that were both important and practical to assess, along with details about the age ranges of assessment and assessment method. Focus was mainly on lifetime prevalence (not current or interval), and need to capture age of onset, age of offset, and number of episodes. A set of impulse control disorders of clear public health

relevance are also assessed (e.g., pathological gambling). Important disorders such as schizophrenia, bipolar disorder, and autism are poorly assessed by the web-based format around which LifeGene will be based, so these will be captured via national registers (i.e., the National Patient Register and the Prescribed Drug Register).

The most common cause of hearing loss is genetic (50-70 %). Studies have shown that carriers of gene deficits also influence the sensitivity to other risk factors, such as occupational exposures. Hearing loss, and its progression, can be caused and/or modified by many different external (e.g. exposure at work, day care centres and free time) as well as internal risk factors like genetic and life style factors. The prevalence of hearing loss has not been investigated, using audiometric measurements, in a large sample of the general population. Hearing tests in LifeGene would provide robust knowledge whether the increasing prevalence of hearing loss is true and also validate the self-estimated hearing. Inclusion of children and young people in LifeGene would also help answering the question whether hearing loss in the "MP3/I-pod" generation is increasing and, if so, how this is influenced by age and other factors. A prospective study design would also give knowledge of the progression of hearing loss over time, both in the population and in individuals (16).

Physical Activity, Nutrition and Dietary Assessment

The broad objectives of the Physical Activity, Nutrition and Dietary Assessment (PANDA) working group were to investigate the feasibility, validity and reliability of a range of methods for the assessment of diet and physical activity suitable for LifeGene. The results from new (LifeGene sponsored) and existing studies were summarised, and recommendations on how the measurement instruments might be applied in LifeGene were provided. More detail is provided in the pre-pilot section, page 22.

E-epidemiology

E-epidemiology is the science underlying the acquisition, maintenance and application of epidemiological knowledge and information using digital media such as the Internet, mobile phones,

digital paper, and digital TV. E-epidemiology also refers to the large-scale science that will increasingly be conducted through distributed global collaborations enabled by the Internet (17). A critical condition in introducing the Internet and other communication technologies in population-based studies is the degree of access to the techniques among the population. In 2008, up to 84% of the Swedish population used the Internet regularly according to Statistics Sweden. For all age groups, using these tools is more common among men than women, but these differences are decreasing year to year. A web-based application for collection of data is easy and inexpensive to construct and maintain (18-20). Web technologies also bring possibilities for enhanced data collection not possible through the traditional approaches, including real-time data collection, interactivity, tailored and personalised questionnaires and repeated measures. Hence, a key feature of LifeGene will be exploitation of modern digital tools.

Sample design and participants

Baseline study

Based on the working groups' recommendations, suggested sample design and recruitment of participants was set up. In the sample design finally selected, index persons (aged 18-45 years) will be randomly sampled from the general population, with oversampling of twins from the Swedish twin registry who have recently been screened with an instrument similar to the LifeGene questionnaire. Participants will be invited to include their household members (partner and any children). By collecting information on households, shared environmental risks and exposures can be quantified. Recruiting children living in the shared household introduces information on persons who share environment plus varying degrees of relatedness, as well as contribute to the life course information 0-45 years. There will also be provision for spontaneous sign-up, to improve power and include participants in a cost-efficient way.

The planned cohort size of 500,000 participants is the same as recommended by an expert panel assembled by the US National Human Genome Research Institute in 2004 (21). For a moderately rare

disease (yearly incidence of 50-100 cases per 100,000) and a follow-up of five years, this is expected to generate a sufficient number of cases to detect a gene-environment interaction with an odds-ratio of 2.5-3 with 80% probability at a conservative significance level of α=0.0001 (to protect against false positives). This holds under fairly conventional assumptions about allele frequencies, mode of inheritance and prevalence of environmental exposures (22).

Annual follow-ups

The study participations will be prompted annually to respond to a short, web-based questionnaire for updates on changes in household composition, changes in symptoms, injuries and pregnancy.

Event-based sampling

The LifeGene study design allows event-based sampling of specific populations. In the early phases of LifeGene, three types of event-based sampling will be included.

1) The LifeGene-Influensa Like Illness (ILI) study addressing influenza in event-based sampling from volunteer participants was launched during the LifeGene pilot in September 2009. The LifeGene infrastructure and possibilities to sample in response to events was planned before the H1N1 viral epidemic emerged. Event-based reports in the LifeGene-ILI allowed sampling prior to disease onset (and vaccination), a household sampling triggered by a positive H1N1 or Corona virus diagnosis confirmed by PCR and followed by a post-seasonal sampling. DNA, serum and viable peripheral mononuclear cells were stored for further analysis. Other tests gauge the immune response directed against the current H1N1 strains as well as previous influenza pathogens.

2) "Born into Cohort": The LifeGene study design includes recruitment of individuals who may become pregnant during the first 6 years of baseline data collection. Thus, LifeGene plans to contact pregnant women (and their partners) with additional web-based questionnaire and biosamples during pregnancy and delivery. With a target of 500,000 individuals in the cohort, a simulation indicates that approximately 21,000 children will be born into the cohort and followed prospectively.

3) Injuries: During the annual follow-ups of LifeGene, participants will be asked about injuries during the past year. Positive response to these items will generate a more in-depth assessment of causes and consequences of these injuries.

Measures and study procedures

Based on the recommendations of the working groups, LifeGene was designed to include collection of data through web-based questionnaires and in person testing to collect biosamples and a range of physical tests, Figure 1. In preparation for the LifeGene pilot, a protocol for web-based questionnaires and in-person testing (biosamples and physical examinations) was developed. Names and addresses of potential index participants will be randomly sampled from the general population, and an invitation letter with personal log-in information sent out. After agreement and consent on the LifeGene webpage, they will be able to respond to the LifeGene survey, book an appointment at the test centre for in-person testing and invite household members to participate. After the first invitation letter, up to three reminders will be sent out. Questions that potential participants may have can be directed to the toll-free number of the LifeGene Support Centre, or via the internet.

Questionnaires

The LifeGene core survey is designed to collect information about the physical, mental and social well-being in the Swedish population, and LifeGene therefore collects data on the diseases and disorders most common in Sweden. Extensive sections of the survey are devoted to assessment of exposures that may be relevant for these outcomes. All survey data are collected through a web portal.

The LifeGene survey is based on a library that distributes relevant questions to the LifeGene participants through a web portal. There are three main categories of study participants entering the web portal: adults aged 18 to 45 years (index persons) or older, children invited by index persons and the parents to these children. The adult library holds approximately 1,350 questions and the child / parent library approximately 1,150 questions.

The questions are available to the study participants through a web portal, showing a circular clock-like menu with questionnaire themes on the dial. Nine themes are shown to adults: Lifestyle, Self-care, Woman's health, Living habits, Health history, Asthma and allergy, Injuries, Mental health and Sociodemography, and between four to nine themes to the partners and children (Figure 2). Parents answer for their children aged 0 to 14 and children answer for themselves from 11 and up, which means that there are parallel questions to children and parents between the ages of 11 and 14.

In-person testing

At the test centres, the study participants are examined for weight, height, waist, hip and chest circumference, bioimpedance, heart rate and blood pressure, along with audiometry and spirometry, with adaptations in individuals above or below 7 years of age. Blood and urine samples are also taken at the test centre for analysis and bio-banking. Table 1 gives details on in-person testing in the LifeGene pilot. In adults, EDTA whole blood for DNA and front-end chemistry will be obtained, along with tubes of EDTA with gel, citrate, lithium heparin and trace metals. In individuals 7 -18 years old, only tubes with EDTA for DNA, EDTA with gel, lithium heparin, trace metals and urinary samples will be collected, and for those under 7 years only EDTA whole blood for DNA and a urinary sample. All samples will be aliquoted for subsequent processing and storage. DNA will be extracted concurrent with other sample processing initially, to be available for subsequent genotyping. The extent of genotyping and choice of platform will be dictated by researcher interest and methodological developments.

Infrastructure

Biobanking

The biological samples collected during in-person testing will be transported in a refrigerated chain with temperature monitors to a biobank facility where primary sample tubes will be processed and stored. The ISO accredited Biobank at Karolinska Institutet is currently establishing new and cost-effective automated processes to allow for novel high throughput techniques. The input processes include registration, volume measurement, sorting with respect to sample type, centrifugation, de-

capping and aliquoting of multiple 225 µl plasma and urine fractions into 2D bar-coded tubes and DNA extraction from 400 µl EDTA blood. Planned processes also include RNA extraction, cell isolation, and front-end chemistry based on blood and urine. The storage solution consist of three elements: (1) a farm of liquid nitrogen tanks housing tens of millions of samples (2) one or several robotic re-picking stations to retrieve desired samples and (3) in the future a large-scale automatic freezer.

The analysis of samples for research takes place after retrieval from storage. Through stringently standardized sampling protocols and modern biobanking, LifeGene will provide an opportunity to evaluate the full ´omics´ – set including proteomics, metabolomics, genomics and epigenomics. Researchers interested in the LifeGene materials will be offered to use external quality-approved analytical centres. Some analysis platforms in high demand will be set up in-house.

IT structure

For LifeGene IT structure, it will be important to provide a long term platform for data collection and withdrawal. We have developed an IT infrastructure and functionality based on process analysis applying modular, rather than single, integration solution approach, built on proven industrial solutions. Integration has been done using web services based on Microsoft.Net and all research data for collection are stored in a central repository with no sensitive data stored at LifeGene test centres.

Biostatistics

The household based longitudinal design of LifeGene allows great flexibility in the type of research questions that can be addressed. It was emphasised that classical concerns of observational epidemiology, such as bias due to selection, confounding and reverse causation are a reality, as well as issues of incomplete data due to non-compliance, drop-out and measurement errors.

For many studies based on LifeGene it may prove useful also to draw on modern statistical methods for observational data, as discussed by Hernán and Robins (23). They set up a framework for causal modelling, which mimics the rationale for controlled randomised trials, and which can disentangle

complex dependencies between time varying exposures and confounders. Referring to the rationale in Prentice et al (24) for the Women's Health Initiative, it is also worth paying attention to the systematic and random errors in the assessment of exposures, in particular assessment of dietary intake and physical activity, and the implications that these measurement errors may have on the estimates of associations between exposure and outcome.

Special design, quality control and analysis issues arise when dealing with high dimensional multi-omics data. These may provide biomarker profiles for early detection or a basis for disease classification, prognosis and treatment prediction. In particular, metabolomics could meet nutritional epidemiology in the search for novel exposure biomarkers that could calibrate traditional measurements for diet and physical activity. Mendelian randomisation could be a useful tool to infer causal relationship between a modifiable exposure and disease (25).

In large randomised controlled trials, it is common practice to augment the original study design with observational components. Conversely, randomised components of lifestyle interventions could be superimposed on the LifeGene observational cohort (26). This would provide analytic advantages for the study of gene-lifestyle interactions, and such "multi-life" interventions would allow volunteer participants to take more active part in the LifeGene assembly of data.

Ethics synopsis

A document on LifeGene ethics policy including details on recruitment, consent, data and sample access and governance has been developed. Further details on the ethics document can be found at www.lifegene.se/For-Scientists1/Ethics/ . Permission for the pilot study was obtained from the Regional Ethical Review Board of Stockholm, Sweden. Researchers requesting to use the data and/or samples will be required to apply for Ethics approval. An impartial data access committee will vet requests for access to de-identified data and samples.

Results

In preparation for the pilot, a number of pre-pilots were performed.

Pre-pilots

1) Physical activity and dietary assessment methods were validated in pregnant and non-pregnant women (aged 20-35 yrs). A physical activity assessment feasibility study was also conducted in young mothers and their four month old babies. The studies compared estimates of physical activity energy expenditure derived from three activity monitors with validation obtained from the standard criterion method of doubly labelled water (DLW).

2) A meal-based food frequency questionnaire (Meal-Q) was developed for the web-based questionnaire and validated against DLW, nutrient-related biomarkers and/or seven days of food records in different age groups. Due to limitations on numbers of food items that can be assessed in a short time, it is important to adapt the method to diets currently consumed by adults, teens, adolescents, children and infants. The selection of food items to be included in Meal-Q was therefore based on age-specific studies using 24 hour recall or food records. Physical activity questionnaires (Active-Q) were developed that were similarly adapted to the usual domains of leisure and occupational activity as well as regular transportations in the various age groups, and validated using DLW, accelerometers and/or seven days records of activities in different age groups.

3) In another LifeGene pre-pilot, dust samples were collected to assess the stability of allergens and endotoxins over time and the variability in allergen content related to fresh, frozen and thawed extracts.

LifeGene Pilot

Based on the recommendations of the working groups, we performed a full “dress-rehearsal” pilot of 1% of the target population, i.e. 5,000 tested individuals. Based on the Swedish population, the

proposed pilot size allows us to estimate the response rate in each age and sex stratum within 5% with 95% probability. The testing sites were chosen to reflect various combinations of geographical location, population density, and association with a university. Thus, we chose a large city (Stockholm, capital of Sweden with almost 2 million inhabitants), a university city (Umeå, in the north with 111,000 inhabitants) and a smaller town (Alingsås, in the west with 36,000 inhabitants). The LifeGene pilot study was launched in Stockholm in October, 2009, followed by the Umeå pilot in late November 2009 and Alingsås in January 2010, and all testing was stopped on March 31, 2010. The testing schedule was designed to generate the number of participants at the proposed testing centres that would stretch the limits of the biobanking facility's ability to process samples within specified parameters and fully test logistics regarding chilled transport. Figure 3 shows the expected age distribution of participants in the pilot.

Influenza pilot

The LifeGene study design allows event-based sampling. An exercise in this context is the LifeGene pilot addressing 'Influenza-like-illnesses', ILI. This study was planned before the H1N1 viral epidemic emerged. The LifeGene-ILI pilot study included sampling of 2,000 adults prior to disease onset (and vaccination), mechanisms for reporting influenza-like symptoms and providing nasal swabs for viral analysis, household sampling triggered by a positive H1N1 diagnosis confirmed by PCR and followed by a post-seasonal sampling. DNA, serum, plasma and viable peripheral mononuclear cells are stored for further analysis; other tests gauge the immune response directed against the current H1N1 strains, as well previous influenza pathogens.

The pre-pilots were concluded during the spring of 2009, the actual pilot was stopped by design at the end of March, 2010 and is currently being evaluated. Recruitment for the main study will start November, 2010 with a gradual expansion of testing centres in the remaining four cities with major universities followed by selected locations in smaller towns, and is planned to take six years.

Discussion

LifeGene is a prospective cohort study with the aim to combine advances in modern biotechnology and information on individuals' health and lifestyle. A comprehensive baseline questionnaire designed to accommodate cutting-edge research questions, state of the art biosampling, repeated follow-ups including event-based sampling are key features of this ambitious initiative. The wide array of high-tech tools encompassed by e-epidemiology will facilitate the longitudinal aspects of the study. Recruitment of index people 18-45 years old, who are invited to include their adult household members and children, will increase the opportunity to involve young couples prior to and during pregnancy, allowing for the first study of children born into a cohort with complete pre- and perinatal data from both mother and father. Furthermore, by focusing on index individuals in young and mid-adulthood in a household structure, we will be able to assess exposures and disease progression for a myriad of disorders that have considerable consequences not only for health care demand, but also for future development of chronic diseases.

Large population-based studies with repeated collection of questionnaire data and sampling over time are extremely costly and the funding has to be well invested. Given the numerous targets that have already been identified through other studies throughout the world, care has to be taken when designing and planning yet another large-scale study. Future studies have to consider carefully whether it is most cost-effective to monitor large populations in prospective cohort studies or undertake well-designed case-control studies on a particular exposure or disease (21). It has been argued that population-based cohorts are necessary to account for methodological challenges such as standardisation and generalisation, to allow for new methodologies and to include younger age groups (27). Others suggest that initiating new large population-based cohorts are not worth the wait and should be replaced by merging ongoing smaller cohort studies (28). In this context, LifeGene has been designed to address a wide variety of research questions, including the possibility of detecting pre-clinical markers for disease, event-based sampling and a cohort of children born into the study.

LifeGene – population-based cohort

We would also like to take this opportunity to invite the scientific community to participate in LifeGene, by proposing sub-studies and possible add-ons. The LifeGene project is a national population-based study of global importance.

Acknowledgment

The authors would like to gratefully acknowledge all the scientists who have contributed with their time and expertise, without which this study would not have been possible to prepare. Their names and affiliations are listed on the LifeGene website: www.LifeGene.se. Start-up funding has been received from Karolinska Institutet, the Stockholm County Council and the Swedish Research Council. Funding has also been obtained from the Torsten and Ragnar Söderbergs Foundation and AFA Försäkringar.

Website

www.LifeGene.se

References

1. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics. 2005 Sep;6(6):639-46.
2. Gulcher J, Stefansson K. Population genomics: laying the groundwork for genetic disease modeling and targeting. Clin Chem Lab Med. 1998 Aug;36(8):523-7.
3. Nakamura Y. The BioBank Japan Project. Clin Adv Hematol Oncol. 2007 Sep;5(9):696-7.
4. Chen Z, Lee L, Chen J, Collins R, Wu F, Guo Y, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). Int J Epidemiol. 2005 Dec;34(6):1243-9.
5. Naess O, Sogaard AJ, Arnesen E, Beckstrom AC, Bjertness E, Engeland A, et al. Cohort profile: cohort of Norway (CONOR). Int J Epidemiol. 2008 Jun;37(3):481-5.
6. Stolk RP, Rosmalen JG, Postma DS, de Boer RA, Navis G, Slaets JP, et al. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. Eur J Epidemiol. 2008;23(1):67-74.
7. Nilsen RM, Vollset SE, Gjessing HK, Skjaerven R, Melve KK, Schreuder P, et al. Self-selection and bias in a large prospective pregnancy cohort in Norway. Paediatr Perinat Epidemiol. 2009 Nov;23(6):597-608.
8. Golding J, Pembrey M, Jones R. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. Paediatr Perinat Epidemiol. 2001 Jan;15(1):74-87.
9. Landrigan PJ, Trasande L, Thorpe LE, Gwynn C, Lioy PJ, D'Alton ME, et al. The National Children's Study: a 21-year prospective study of 100,000 American children. Pediatrics. 2006 Nov;118(5):2173-86.
10. Keil T, Kulig M, Simpson A, Custovic A, Wickman M, Kull I, et al. European birth cohort studies on asthma and atopic diseases: II. Comparison of outcomes and exposures--a GA2LEN initiative. Allergy. 2006 Sep;61(9):1104-11.
11. Inskip HM, Godfrey KM, Robinson SM, Law CM, Barker DJ, Cooper C. Cohort profile: The Southampton Women's Survey. Int J Epidemiol. 2006 Feb;35(1):42-8.
12. Lombardi VC, Ruscetti FW, Das Gupta J, Pfost MA, Hagen KS, Peterson DL, et al. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. Science. 2009 Oct 23;326(5952):585-9.
13. Barker DJ. Fetal origins of coronary heart disease. Bmj. 1995 Jul 15;311(6998):171-4.
14. Brooks PM. The burden of musculoskeletal disease--a global perspective. Clin Rheumatol. 2006 Nov;25(6):778-81.
15. The global burden of disease; 2004 update. Available from www.who.int/evidence/bod. Geneva, Switzerland: WHO Press; 2006 [cited 2010, June 14].
16. Shargorodsky J, Curhan SG, Curhan GC, Eavey R. Change in prevalence of hearing loss in US adolescents. JAMA. 2010 Aug 18;304(7):772-8.
17. Ekman A, Litton JE. New times, new needs; e-epidemiology. Eur J Epidemiol. 2007;22(5):285-92.
18. Ekman A, Dickman PW, Klint A, Weiderpass E, Litton JE. Feasibility of using web-based questionnaires in large population-based epidemiological studies. Eur J Epidemiol. 2006;21(2):103-11.
19. Ekman A, Klint A, Dickman PW, Adami HO, Litton JE. Optimizing the design of web-based questionnaires--experience from a population-based study among 50,000 women. Eur J Epidemiol. 2007;22(5):293-300.
20. Bexelius C, Honeth L, Ekman A, Eriksson M, Sandin S, Bagger-Sjoback D, et al. Evaluation of an internet-based hearing test--comparison with established methods for detection of hearing loss. J Med Internet Res. 2008;10(4):e32.
21. Collins FS. The case for a US prospective cohort study of genes and environment. Nature. 2004 May 27;429(6990):475-7.

22. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009 Feb;38(1):263-73.
23. Hernán MA, Robins JM. Observational studies analyzed like randomized experiments: Best of both worlds. Epidemiology. 2008;19:789-92.
24. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. Biometrics. 2005 Dec;61(4):899-911; discussion -41.
25. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003 Feb;32(1):1-22.
26. Ioannidis JP, Adami HO. Nested randomized trials in large cohorts and biobanks: studying the health effects of lifestyle factors. Epidemiology. 2008 Jan;19(1):75-82.
27. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. Nature. 2007 Jan 18;445(7125):259.
28. Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: not worth the wait. Nature. 2007 Jan 18;445(7125):257-8.

Tables and Figures

Figure 1. Flow chart on the structure of the study. Questionnaires (Q), parental questionnaires (pQ) and In Person Testing (IPT) in adults, adolescents and children.

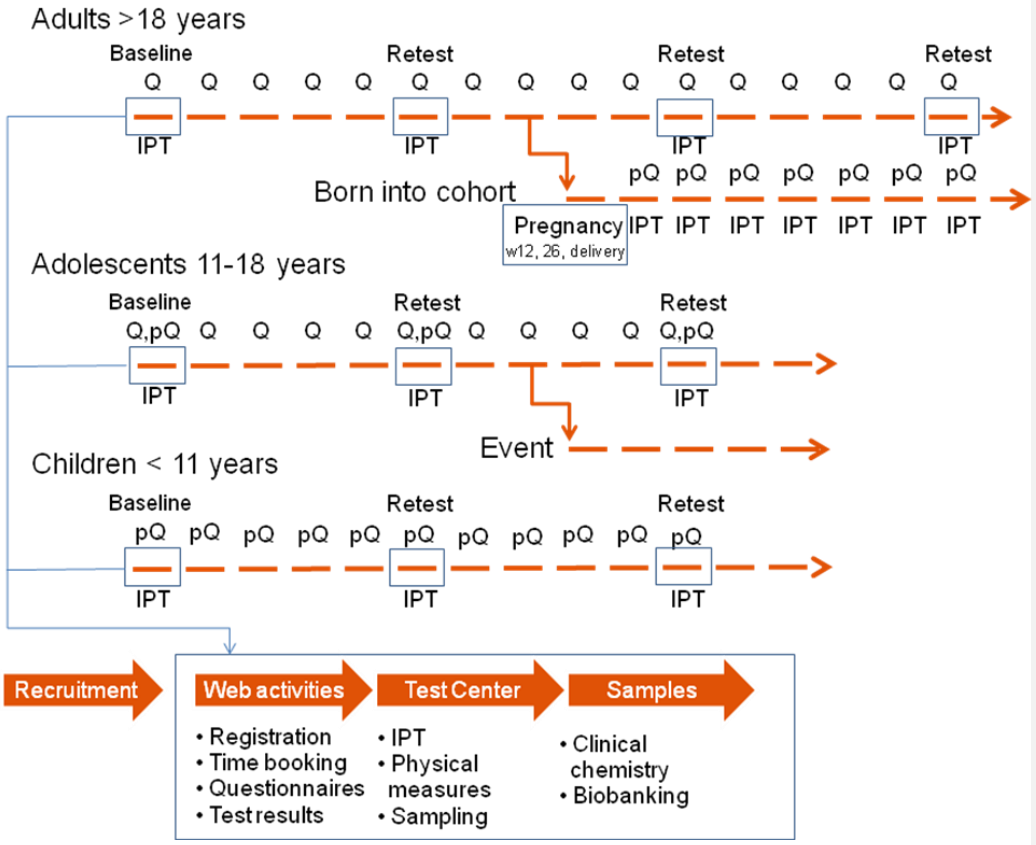
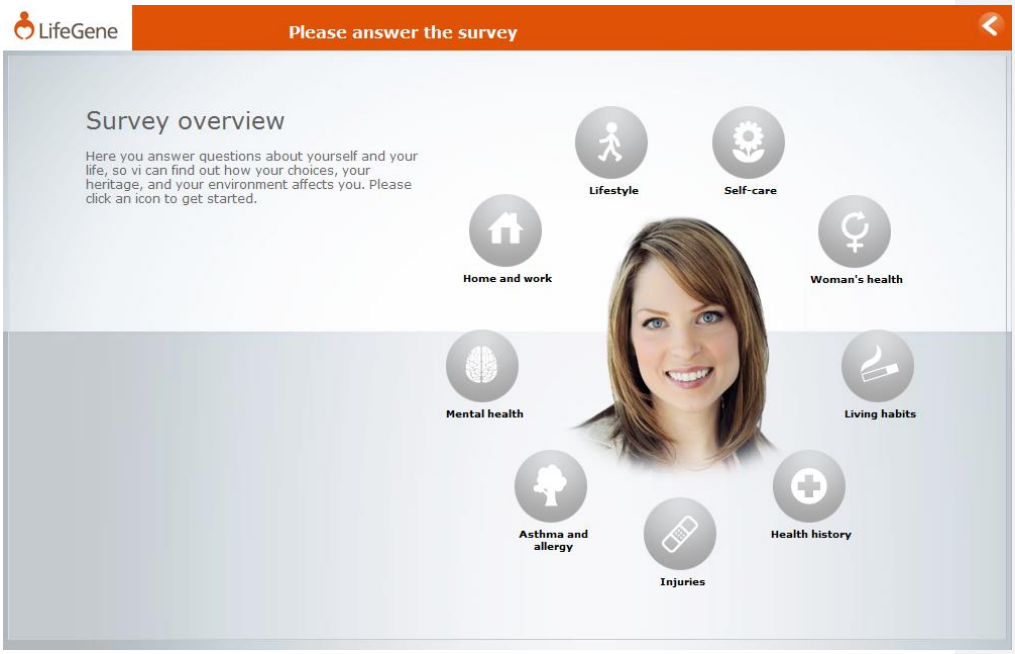


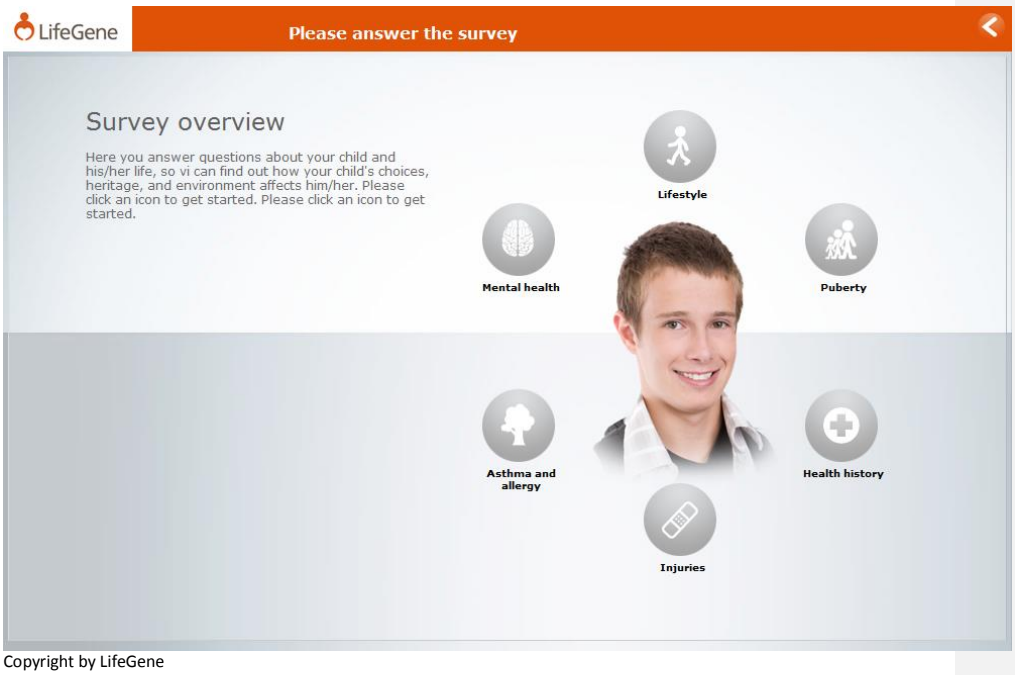
Figure 2. Questionnaire modules in adults (a) and children (b)

a)



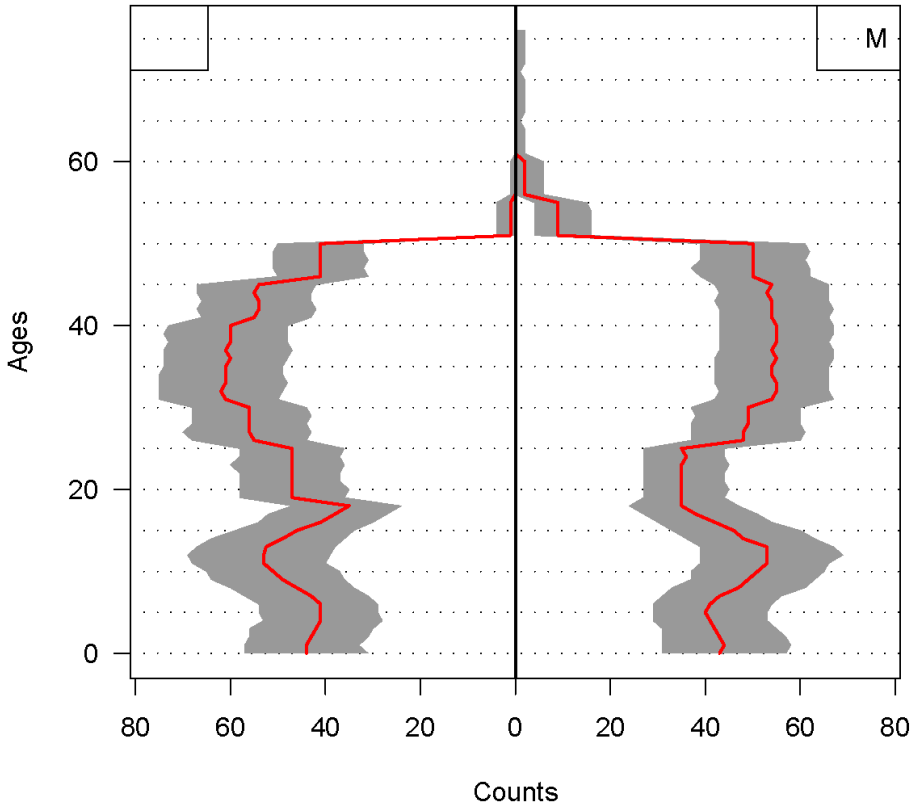
Copyright by LifeGene

b)



Copyright by LifeGene

Figure 3. Population distribution of the LifeGene pilot at the end of the recruitment period. Red line: median over 100 simulations, dark grey are: 95% for 100 simulations.



LifeGene – population-based cohort

Table 1. Biosamples and physical measurements collected during the pilot

Biosamples	**Purpose**	**Adults**	**7-18 yrs**	**<7 yrs**
EDTA whole blood	DNA	x	x	x
EDTA whole blood	WBC+diff/HbA1c	x		
LiHep with gel	Frontend chemistry	x		
EDTA with gel	Aliquots	x	x	
EDTA with gel	Aliquots	x		
Citrate 3.8%	Aliquots	x		
LiHep with gel	Aliquots	x	x	
Trace metal	Aliquots	x	x	x
Urine	Aliquots	x	x	x

Physical measures	**>7 yrs**	**6 yrs**	**5 yrs**	**<5 yrs**
Height	x	x	x	x
Weight	x	x	x	x
Bioimpedance	x			
Waist circumference	x	x	x	x
Hip circumference	x	x	x	x
Thorax circumference	x	x		
Heart rate	x	x	x	
Blood pressure	x	x	x	
Audiometry	x	x		
Spirometry	x	x		