International Journal of Epidemiology, 2015, 1137–1147
doi: 10.1093/ije/dyt268
Advance Access Publication Date: 11 February 2014
Cohort Profile

Cohort Profile

Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu

Liis Leitsalu,[1,2] Toomas Haller,[1] Tõnu Esko,[1,3,4,5] Mari-Liis Tammesoo,[1] Helene Alavere,[1] Harold Snieder,[1,6] Markus Perola,[1,7,8] Pauline C Ng,[1,9] Reedik Mägi,[1] Lili Milani,[1] Krista Fischer[1] and Andres Metspalu[1,2,10]*

[1]Estonian Genome Center, University of Tartu, Tartu, Estonia, [2]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [3]Divisions of Endocrinology, Boston Children's Hospital, Boston, MA, USA, [4]Department of Genetics, Harvard Medical School, Boston, MA, USA, [5]Broad Institute of Harvard and MIT, Cambridge, MA, US, [6]Department of Epidemiology, University of Groningen, Groningen, The Netherlands, [7]Public Health Genomics Unit, Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland, [8]University of Helsinki, Institute for Molecular Medicine, Helsinki, Finland, [9]Genome Institute of Singapore, Singapore and [10]Estonian Biocentre, Tartu, Estonia

*Corresponding author. Estonian Genome Center,University of Tartu 23b Riia St. Tartu 51010, Estonia.
E-mail: andres.metspalu@ut.ee

Accepted 3 December 2013

Abstract

The Estonian Biobank cohort is a volunteer-based sample of the Estonian resident adult population (aged ≥18 years). The current number of participants—close to 52000-—represents a large proportion, 5%, of the Estonian adult population, making it ideally suited to population-based studies. General practitioners (GPs) and medical personnel in the special recruitment offices have recruited participants throughout the country. At baseline, the GPs performed a standardized health examination of the participants, who also donated blood samples for DNA, white blood cells and plasma tests and filled out a 16-module questionnaire on health-related topics such as lifestyle, diet and clinical diagnoses described in WHO ICD-10. A significant part of the cohort has whole genome sequencing (100), genome-wide single nucleotide polymorphism (SNP) array data (20 000) and/or NMR metabolome data (11 000) available (http://www.geenivaramu.ee/for-scientists/data-release/). The data are continuously updated through periodical linking to national electronic databases and registries. A part of the cohort has been re-contacted for follow-up purposes and resampling, and targeted invitations are possible for specific purposes, for example people with a specific diagnosis. The Estonian Genome Center of the University of Tartu is actively collaborating with many universities, research institutes and consortia and encourages fellow scientists worldwide to co-initiate new academic or industrial joint projects with us.

Key Messages

- The Estonian Genome Center of the University of Tartu belongs to international networks such as BBMRI (Biobanking and Biomolecular Resources Research Infrastructure), BBMRI-ERIC (European Research Infrastructure Consortium) and the Public Population Project in Genomics (P^3G).
- A wide range of phenotypes have been investigated, with over 200 traits and sub-phenotypes under analysis involving anthropometric traits and blood biochemistry, metabolomics, common and rare diseases, personality and lifestyle.
- Core funding for the Estonian Biobank is provided by the Estonian Government.
- The Human Genes Research Act was created in 2000 to ensure the legal basis for the Estonian Biobank.

Why was the Biobank initiated?

The Estonian Genome Project Foundation initiated the Estonian Biobank Project in 1999, which was transformed into the Estonian Genome Center of the University of Tartu (EGCUT) in 2007.[1,2] The objective was to investigate the genetic, environmental and behavioural background of common diseases and traits by creating a biobank with biological samples and health records from a large proportion of the population. Since 2012, the biobank has been sited in a new building specifically designed for the EGCUT and biobanking purposes in Tartu, Estonia.

To enable the establishment and ensure the legal basis for the Estonian Biobank, the Estonian Human Genes Research Act (HGRA)[3] was passed by the Parliament of Estonia in 2000. According to this law, the EGCUT has the right to use and store personal data and biological samples for three purposes:

i. to promote the development of genetic research;
ii. to collect information on the health status of the Estonian population combined with genetic information;
iii. to use the results of genetic research to improve public health.

Investigating the population's health and '-omics' data will be accomplished through a longitudinal population-based cohort study of a large variety of samples with extensive clinical information. The information is continuously updated through follow-up on health status using both electronic population registries and re-examinations of the participants.

About €12 million from both the private and public sectors has been spent between 2000 and 2013 to establish the Estonian Biobank with the health records, DNA, plasma and white blood cell (WBC) samples from 52 000 participants. Since 2007, the core funding for the Estonian Biobank at the EGCUT is provided by the Estonian Government through the budgets of the Ministry of Social Affairs, the Ministry of Economics and Communications and the Ministry of Education and Research. The financing of the EGCUT and the Estonian Biobank is regulated through the HGRA. Funding for the research projects is competitively based. About €10 million has been received for infrastructure development and research projects.

Who is in the cohort?

The Estonian Biobank has reached a cohort size of 52 000 participants aged 18 years and older.[4] The age, sex and geographical distribution closely reflect those of the Estonian adult population (see Figures 1 and 2) and encompass close to 5% of the entire adult population of Estonia,[5] making the Estonian Biobank internationally

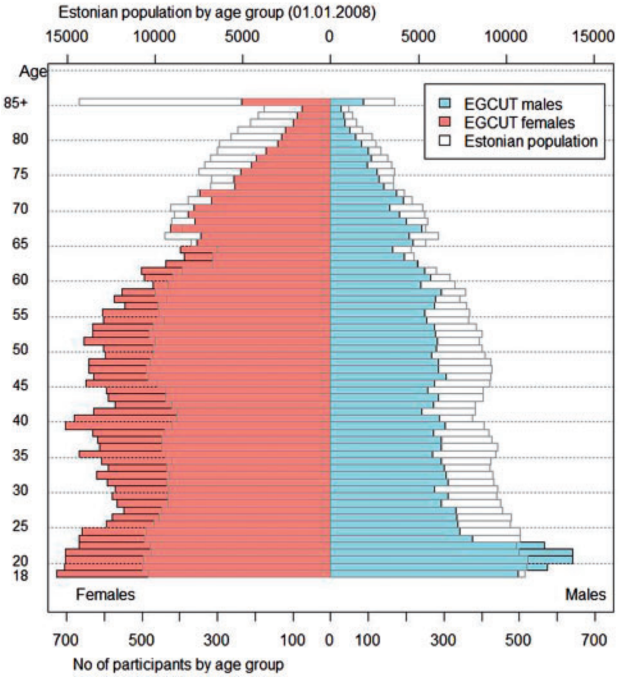

Figure 1. Age in years and gender distribution of the participants at recruitment in comparison with the adult population of Estonia. Number of participants is considered as of March 2011 with a total of 50 916 participants. Estonian population is estimated as of 1 January 2008 by Statistics Estonia. The age groups above 85 years have all been merged into one group of 85+.

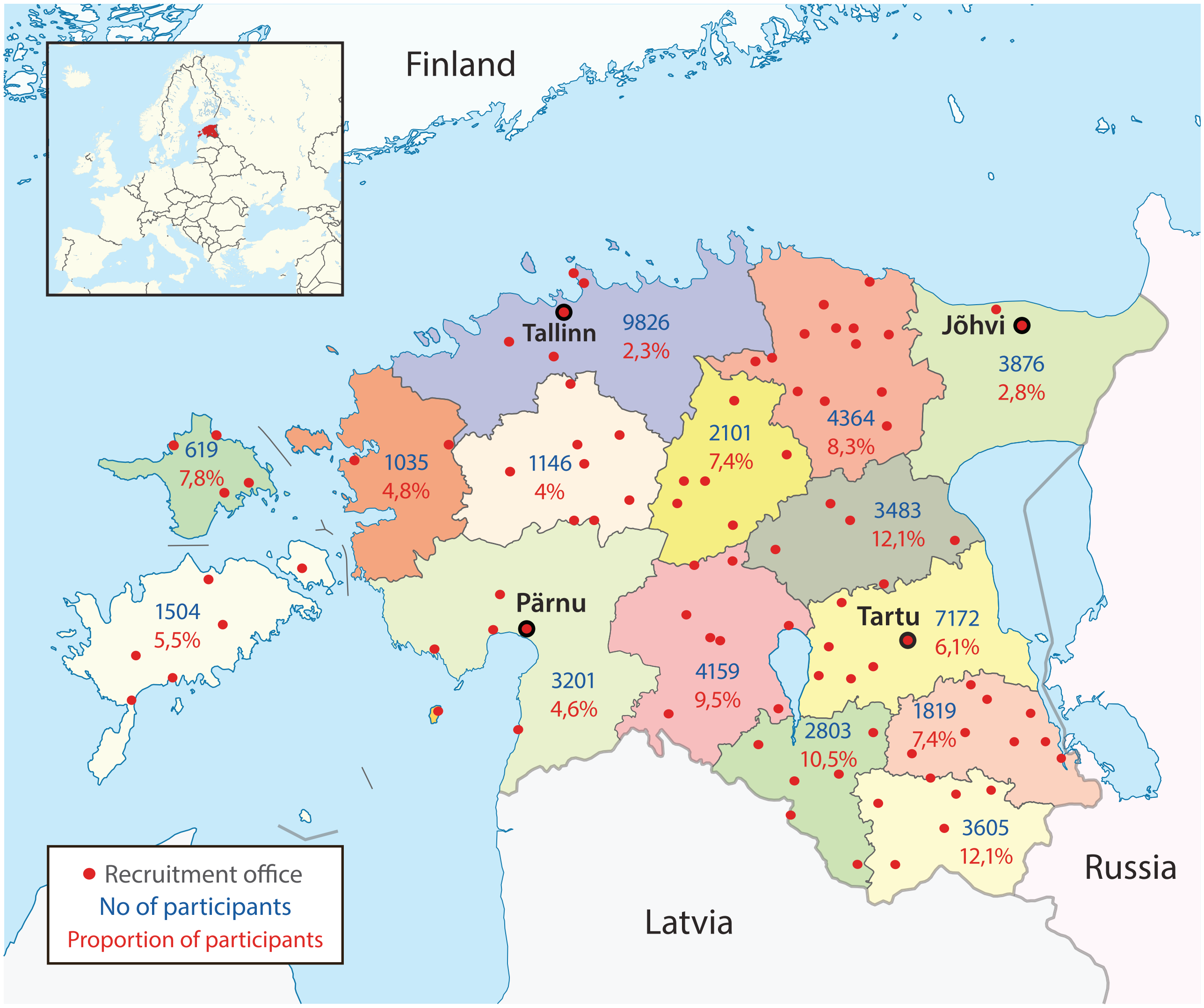

Figure 2. Map of Estonia showing the 15 counties, the network of recruitment offices and the proportions (%) of participants derived from each county. The total number of participants is considered as 50 916. Since 2012, only two recruitment offices remain in Tartu.

competitive in terms of the absolute number of participants as well as the relative number of participants as a proportion of the entire population—all being characterized according to the same protocol.

When comparing the numbers of women and of men in the EGCUT database with those in the Estonian population, one can observe that as in most other voluntary population-based studies, females participate more actively than males (Figure 1).[6–8] There are proportionally fewer males in almost all age groups except for the group aged 18–21 years; women are represented in large numbers in all age groups up to 60 years. Overall, the representation of men in the biobank is 3.4% and of women is 5.5% of the adult population of Estonia, and the representation of men and women in the biobank is 34% and 66%, respectively, compared with 45% and 55%, respectively, in the general population. Older people tend to participate less frequently; however, all age groups are well represented. The gaps in our population caused by the Second World War, for instance, are reflected by the gaps observed in the age groups of 62–66 and 40–44 years.

What has been measured and collected?

For the recruitment of participants and for the collection of samples and health data, a unique network of data collectors was set up consisting of GPs and other medical personnel in private practices and hospitals or in the recruitment offices of the EGCUT (Figure 2). Geographically, all 15 Estonian counties participated and a total of 454 GPs, representing around 56% of all registered GPs,[9] and 186 nurses were involved.

The questionnaire was initially developed in 2001 in collaboration with the World Health Organization's International Agency for Research on Cancer (IARC) in Lyon, France (Prof. Elio Riboli). The list of data items and

Table 1. Data items collected by the EGCUT for the Estonian Biobank

Year	Measurements and samples
2002–10	Baseline questionnaire and measurements: • Computer-assisted genetic epidemiological questionnaire is filled. Information available in the electronic medical records is added to self-reported health information while the source of the information is recorded • Anthropometric measurements taken: weight, height, waist and hip circumferences, hair and eye colour, dominant hand • Blood pressure and resting heart rate are measured • Venous blood sample (30–50 ml) is drawn for DNA, WBCs and plasma
2003	Extra modules were added to the baseline questionnaire for participants with hypertension
2004	Extra modules were added to the baseline questionnaire for participants with type 2 diabetes.
2007	• Sleep module (MCTQ) and psychiatric module (MINI and SSP) were added to the questionnaire (for participants with specific diagnoses only) • Personality test (NEO-PI-R) added to the questionnaire with a separate informed consent form
2011–12	Food neophobia module[10] added to the baseline questionnaire Additional samples taken as part of a follow-up project: • Buccal swabs • Venous blood sample for RNA and serum for clinical biochemical analysis
2012	Dynamometrics, electrocardiogram, and spirometry were added to the measurements taken by the recruiter as part of a follow-up project. All equipment was standardized and the standard operating procedures were harmonized with the German National Cohort[11]

MCTQ, Munich Chronotype Questionnaire;[12] MINI, Mini-International Neuropsychiatric Interview;[13] SSP, Swedish universities Scales of Personality;[14] NEO-PI-R, Neuroticism-Extraversion-Openness Personality Inventory Revised.[15]

samples collected has grown in the past 12 years (Table 1). The latest version of the questionnaire (Figure S1, available as Supplementary data at *IJE* online) consists of approximately 330 questions with more than 1000 data fields.[4] To ensure standardized and accurate data collection procedures, a computer-assisted personal interview (CAPI) was used. The software allows the recruiter to fill out the questionnaire together with the participant. The questions were asked in a specific order with a slight variation depending on the age, gender and reported diagnoses of the participant. When obligatory fields were not properly filled in, the recruiter was notified.

The involvement of GPs as recruiters for the EGCUT has provided several advantages in terms of health data collection. Due to the developments in the Estonian information technology infrastructure, there are electronic health records available for GPs that include pre-existing medical data for the patients. Permitted by the HGRA, this pre-existing medical information is included in the EGCUT phenotypic data of the participants. The GPs also recorded whether the diagnoses were confirmed by a clinician orgeneral practitioner or were self-reported (Table 2). This provides confidence in the reliability of the diagnosis.

For all starting data collectors, the first 10 questionnaires were monitored for completeness and illogical answers, and thereafter 10% of all the questionnaires were selected randomly for monitoring; 21% of the questionnaires have been inspected and corrected when necessary. From the monitored questionnaires, 99% were classified as of high quality, meaning that all the fields were filled in and the answers appeared logical.

The collected samples were transported to the central laboratory of the EGCUT within 24–48 h and every step was recorded with a time stamp. Upon arrival at the central laboratory, the samples were assigned a new 16-digit code by the coding centre to maintain the anonymity of the donors, and each tissue sample was barcode-labelled. The coding centre is a secured area where all phenotype and genealogy data and the keys between the codes for the different information systems are stored in a highly secure environment. The server in the coding centre is not connected to the internet. Data are transferred by USB memory devices. Four separate codes are implemented: a biological sample and data transport code; a unique participant ID; a laboratory code for the biological sample; and a unique release code for data release for further research projects. All data transfers and databases are encrypted.

The DNA, plasma and WBCs are isolated immediately, packaged into CryoBioSystem high security straws (DNA in 10–14, plasma in 7, WBCs in 2 straws) and stored in liquid nitrogen. For intermediate storage of normalized DNA samples, the Hamilton Robotics Automated Sample Management (ASM) system with 100 000 tube capacity is used. The ASM allows quick and accurate selection of samples. All procedures are in accordance with the ISO 9001:2008[16]standard and a custom-made Laboratory Information Management System (LIMS). The LIMS supports the laboratory work processes, quality assessment

Table 2. Selection of ICD-10 diagnoses reported with respective class of reliability

ICD-10 code, diagnosis		Classes of reliability					Number of diagnoses in the database
		1	2	3	4	Unknown[a]	
C50	Malignant neoplasm of breast	110	37	4	82	49	282
C53	Malignant neoplasm of cervix	37	16	2	63	33	151
C61	Malignant neoplasm of prostate	102	25	2	40	20	189
E05	Thyrotoxicosis (hyperthyroidism)	329	57	14	281	100	781
E10	Insulin-dependent diabetes mellitus	222	75	5	108	48	458
E11	Non-insulin-dependent diabetes mellitus	1432	209	15	369	213	2238
E66	Obesity	987	270	152	538	38	1985
E78	Disorders of lipoprotein metabolism and other lipidaemias	2013	157	19	385	42	2616
F20	Schizophrenia	238	34	2	33	4	311
F32	Depressive episode	1176	276	82	1844	413	3791
F33	Recurrent depressive disorder	410	149	24	224	99	906
G20	Parkinson's disease	87	21	3	21	1	133
G40	Epilepsy	270	76	13	184	42	585
I10	Essential (primary) hypertension	3748	554	161	2771	164	7398
I11	Hypertensive heart disease	4321	710	39	640	36	5746
I21	Acute myocardial infarction	644	57	5	185	48	939
J45	Asthma	834	218	29	796	109	1986
K25	Gastric ulcer	487	80	9	784	182	1542
K26	Duodenal ulcer	861	122	31	1096	413	2523
L40	Psoriasis	403	133	34	599	42	1211
M80	Osteoporosis with pathological fracture	102	35	3	57	4	201
M81	Osteoporosis without pathological fracture	304	76	21	559	20	980
N18	Chronic renal failure	199	36	0	12	3	250
N20	Calculus of kidney and ureter	400	64	14	407	187	1072
Total		19 716	3 487	6 83	12 078	2 310	38 274

Number of diagnoses for selected categories as of March, 2011 for 50 916 participants out of a total of 377 871 diagnoses. Classes of reliability: 1 = the diagnoses are confirmed by the family doctor/specialist doctor and the results of the examinations are listed in the family doctor's database; 2 = the diagnoses are proved by a specialist doctor and the results of the examinations are not listed in the family doctor's database; 3 = probable diagnoses (the clinical symptoms and results of the examinations are only partly-proved); 4 = possible diagnoses (diagnoses self-reported by the participant.

[a] The class of reliability of the diagnosis has not been specified.

and control. The completed questionnaires were sent to the EGCUT electronically as an encrypted document, and the consent forms were the only documents that were stored in the coding centre in paper form.

The questionnaire, freely available online, includes personal, genealogical, lifestyle, educational and occupational history questions (Table 3, Supplementary Tables S1 and S2, available as Supplementary data at *IJE* online).[4] Medical history and current health status are recorded in accordance with the International Classification of Diseases (ICD-10) codes,[17] medications according to the Anatomical Therapeutic Chemical (ATC) classification,[18] educational information is recorded according to the International Standard Classification of Education (ISCED)[19] and the occupational information is recorded based on the International Standard Classification of Occupations (ISCO-88).[20]

Altogether there are currently over 377 000 diagnoses for 52 000 participants with approximately 7.4 diagnoses on average per participant (Table S1, available as Supplementary data at *IJE* online). The common diseases are proportionally adequately represented.[21] For example, among the EGCUT participants there are 21 800 individuals with a total of 40 200 diagnoses of the circulatory system. Similarly, there are 7500 participants with 8900 diagnoses of neoplasm in the EGCUT database. Additionally, there are 30 700 participants with 60 800 diagnoses of the respiratory system and 9900 participants with 12 700 diagnoses of endocrine, nutritional and metabolic diseases (Table S1, available as Supplementary data at *IJE* online).

How often have the participants been followed up?

The usefulness of a cohort for research purposes is dependent on the information accompanying the biological samples. The scientific value will decrease unless the collection of information is maintained, constantly updated and, when possible, expanded. The HGRA addresses this issue

Table 3. Anthropometric characteristics measured at baseline, 2003–10, 50 916 participants

Characteristics	Men, %	Women, %	Total, %
Height (cm)			
<150	0.2	1.0	0.7
150-159	0.4	18.8	12.5
160-169	8.7	56.7	40.2
170-179	44.9	22.3	30.1
180-189	39.4	1.2	14.3
>190	6.4	0.0	2.2
Mean (SD)	178.6 (7.1)	164.8 (6.4)	169.5 (9.4)
Weight (kg)			
<50	0.3	2.9	2.0
50-59	2.0	19.8	13.7
60-69	13.2	29.5	23.9
70-79	25.9	21.9	23.3
80-89	25.2	13.6	17.5
90-100	20.0	7.9	12.0
>100	13.5	4.5	7.6
Mean (SD)	84.4 (15.8)	71.2 (15.3)	75.7 (16.7)
Waist circumference (cm)			
<70	19.6	32.1	27.8
70-79	11.8	23.2	19.3
80-89	23.0	18.7	20.2
90-100	24.4	15.2	18.3
>100	21.2	10.8	14.4
Mean (SD)	93.1 (13.7)	84.1 (14.5)	87.2 (14.9)
Hip circumference (cm)			
<70	18.9	20.3	19.8
70-79	0.6	0.5	0.6
80-89	3.6	5.1	4.6
90-100	32.8	30.2	31.1
>100	44.1	43.8	43.9
Mean (SD)	102.2 (9.4)	103.5 (11.7)	103.0 (11.0)
Systolic BP (mmHg)			
<100	0.7	3.1	2.3
100-109	4.2	12.4	9.6
110-119	14.8	22.3	19.7
120-129	28.6	24.9	26.1
130-139	22.7	15.5	17.9
140-149	16.1	11.2	12.9
150-160	9.5	7.5	8.2
>160	3.5	3.1	3.2
Mean (SD)	129.9 (16.3)	124.3 (18.1)	126.2 (17.7)
BMI (kg/m^2)			
Underweight (<20) kkkg)kg/m^2)	1.0	3.2	2.4
Normal (20-24.9)	42.3	45.6	44.4
Overweight (25-29.9)kg/m^2)	36.7	28.0	31.0
Obese (≥30)	19.9	23.3	22.1
Mean (SD)	26.4 (4.7)	26.3 (5.7)	26.3 (5.4)

by permitting the EGCUT to re-contact the participants for follow-up purposes or for collection of additional information not otherwise available. The follow-up data allow scientists to conduct longitudinal studies, making it possible to provide comparisons of epigenomic, metabolomic and phenotypic profiles before and after the onset of a disease.

There have been two projects to re-contact for follow-up purposes. In 2008, a sample of 822 participants, in an age range of 20–74 years, was selected to be re-contacted to validate the initial questionnaire data, as well as to study the dynamics in various phenotypic characteristics and in biological measurements. Re-contacting participants was carried out between 2008 and 2010, with a response rate of 57.2%. A second wave of re-contacting for follow-up purposes started in 2011 and 1072 participants visited our recruitment centres in Tallinn and Tartu for the second time, with a response rate of 41.1%. The difference in response rate could be accounted for different methods of re-contacting. Whereas the first time the participants were contacted through their GP, the second time they were contacted directly by mail. Although the Estonian Biobank updates the participants' home addresses through the Population Register,[22], the addresses in the register do not always correspond to the actual residency of the participant. Hence it is not possible to state the response rate among those who received the invitation.

The HGRA permits the enrichment of phenotypic data by updating the health data from medical records available in national health databases and hospital registries. Linking to registries is conducted semi-annually. So far, the EGCUT has successfully augmented data using the Estonian Population Register, the Estonian Causes of Death Registry, the Estonian Cancer Registry and the Estonian Tuberculosis Registry (Table 4).[23] Based on the data from the Estonian Causes of Death Registry, 2333 participants have died as of August 2013. According to the data received from the Estonian Cancer Registry in June 2012, there are 2538 gene donors with a recorded diagnosis of malignant neoplasm. Based on the data received from the Estonian Tuberculosis Registry, there are 260 participants with a recorded tuberculosis diagnosis. Integration with national health databases allows the EGCUT to periodically retrieve affected status updates on gene donors.

Linking to the databases of the two major hospitals as well as collaboration with the Estonian Health Insurance Fund have also been started.[24] Additionally, linking has been approved by the ethics committee and data protection inspection with the Estonian eHealth Foundation.[25] The future plans include linking to the Myocardial Infarction Registry, after which the Estonian Biobank will be linked to all currently existing national health registries.

What has been found? Key findings and publications

The EGCUT is collaborating scientifically on both national and international levels. One of the first studies established Estonians as a representative population for northern

Table 4 Registries and databases used for periodical data updates for EGCUT gene donors. Baseline information of all participants is regularly updated through linking to the following registries and databases

Registry	Type of information received	Linking conducted
Population Register	Home addresses as provided by the citizens	March 2010, November 2010, October 2011, April 2012, January 2013
Estonian Causes of Death Registry	Date, cause and country of death	September 2010, May 2011, January 2012, December 2012, August 2013
Estonian Cancer Registry	Diagnoses of neoplasms, analyses results	May 2011, June 2012
Estonian Tuberculosis Registry	Diagnoses, therapy	December 2012
Estonian Health Insurance Fund[a] (including Digital Prescription Database)	Diagnoses; prescriptions provided and prescriptions used; service billing information including physician specialty, institution, type of care provided and specialty	December 2012
Database of the Tartu University Hospital[b]	Diagnoses, hospital discharge record, laboratory data, imaging data and results of other analysis	July 2013, August 2013
Database of the North Estonia Medical Centrein Tallinn[b]	Diagnoses, hospital discharge record, laboratory data, imaging data and results of other analyses	January 2013
Linking has also been approved by the ethics committee and data protection inspection:		
Estonian eHealth Foundation	Data such as electronic health records, digital prescriptions from all medical service providers	

[a] Estonian Health Insurance Fund: information should be available on all citizens insured by the national health insurance.

[b] Databases of Tartu University Hospital and the North Estonia Medical Centre in Tallinn: 75% of gene donors have visited or been hospitalized in these two largest hospitals.

Europeans. Using the EGCUT genetic data, Nelis *et al.* investigated the European genetic structure, along with 16 other European populations.[26] This study revealed that the genetic structure of the European populations is well correlated with their geographical locations, and that the Estonian population is not an isolate, like the Finnish or Sardinian populations, but is more heterogeneous.[26,27] On the map constructed by principal component analysis, Estonians clustered closest to Latvians, Lithuanians and northwestern Russians.[26,27]

The EGCUT participates in many phenotype-driven and genetic epidemiological consortia such as GIANT (Genetic Investigation of ANthropometric Traits), CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology), ENGAGE (European Network for Genetic and Genomic Epidemiology), DIAGRAM (DIAbetes Genetics Replication And Meta-analysis), MAGIC (Meta-Analyses of Glucose and Insulin-related traits), CARDIOGRAM (Coronary ARtery DIsease Genome-wide Replication And Meta-analysis) and MAGENETIC (Metabolites And GeNETIcs Consortium). In total, the EGCUT cohort is currently involved in more than 100 international projects. The majority of the collaborators are universities with their own biobank collections, or technology providers such as BGI in Shenzhen, China, and Illumina Inc. from San Diego, CA, USA. The EGCUT participates in several European Union-funded projects such as Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)[28] and BBMRI-Large Prospective Cohorts (BBMRI-LPC).[29] Being one of the founding members of the Public Population Project in Genomics (P^3G),[30] the EGCUT also takes part in this global organization. Due to the EGCUT participation in these consortia, a wide spectrum of '-omics' data are available for a significant part of the cohort (Table 5).

Altogether, a wide range of phenotypes have been investigated, with over 200 traits and sub-phenotypes under analysis: these include anthropometric and blood biochemistry traits,[34–40] common and rare diseases,[41–54] personality and lifestyle,[55–62] and molecular phenotypes[63–70] to highlight some of the scientific findings. One of the motivations of the EGCUT is to investigate these phenotypes to better understand the pathophysiological mechanisms leading to disease onset and come up with biomarkers with better prediction power and accuracy. Such studies would eventually allow moving the paradigm of general healthcare from curing (dealing with disease outcomes) to prevention (changing the lifestyle or preventive drug treatment).

The complete list of publications can be found at http://www.geenivaramu.ee/for-scientists/publications.html.

What are the main strengths and weaknesses?

The main strength of the Estonian Biobank is that it is a population-based biobank with a longitudinal and prospective database. This means that a wide range of age

Table 5. '-omics' data available in the EGCUT database

Method	Sample size
Sequencing[a]	
Whole-genome sequence	100
Exome sequence	1002
Genome-wide genotyping[b]	
IlluminaHumanOmniExpress	10,000
IlluminaHumanCoreExome	5100
IllumnaExomeChip	5000
IlluminaCardioMetaboChip	2700
Illumina HumanHap370CNV	2700
IlluminaImmunoChip	2000
Transcriptome	
Illumina HT12v3 expression arrays	1100
mRNAseq (peripheral blood)	1000
totalRNAseq (peripheral blood)	50
Other '-omics' platforms	
Metabolomics (NMR[c])	11 000
Metabolomics (MS/MS[d])	1100
Telomere length (blood)[e]	5200
Illumina HumanMethylation450 (peripheral blood)	500
Illumina HTv4 expression arrays (CD4+ and CD8+ cells)	300
Illumina HumanMethylation450 (CD4+ and CD8+ cells)	100
Clinical biochemistry	2700

[a] Sequencing: TruSeq sample preparation, 30–80X paired-end sequencing using HiSeq2000 (Illumina).

[b] Some participants have been genotyped with several beadchips.

[c] NMR: measured by nuclear magnetic resonance method.[31]

[d] MS/MS: measured by tandem mass spectrometry.[32]

[e] Measured by quantitative PCR-based technique.[33]

groups and phenotypes are represented. Whereas usually there is a tendency for urban populations to be overrepresented, this is not the case for the Estonian Biobank cohort (Figure 2).

One of the important strengths is that the biobank has available DNA, plasma and WBCs for each donor. This means that it is possible to analyse the direct effects of sequence variants on metabolism. Even more, it is possible to transform the cells into cell lines or induced pluripotent stem (iPS) cells and directly carry out molecular biology or genetics experiments.

Another strength is provided by the HGRA together with the broad consent form that allows participation in a wide range of research projects without having to re-contact and ask for re-consent.[3,71] The HGRA and consent form also permit donors to request the release of their genetic data, hereditary characteristics and genetic risks obtained from genetic research conducted. This would eventually enable the conduct of projects on personal genome testing, risk perception and management in industrial settings.

Additionally, the HGRA allows the Biobank to obtain additional information through linkage of records with the national electronic registries and major hospitals. All the registries are centrally linked through a nationwide technical infrastructure, the so-called X-road, allowing secure data exchange between databases,and the uniform identifiers of individuals enable quick and errorless linking.[72]

The HGRA also imposed restrictions on the activities of the EGCUT and the data collected in the Estonian Biobank. Participation had to be entirely voluntary—only the individuals interested in joining the Estonian Biobank, after hearing about it either at special promotion events, from the media, from friends or at the family physician's office or hospitals, are recruited. The EGCUT was not allowed to send the invitation letters to their home addresses.[3,4] Therefore, the biobank does not represent a classical random sample and could be subject to recruitment bias. A considerable proportion of the population recruited, however, could compensate for this bias. Hence, although not classically random, the cohort can still be considered representative of the population. Although recruitment was open to everyone, there is a disproportion of ethnic Estonians and ethnic Russians in the biobank, with Estonians being overrepresented (81% in the biobank compared with the 70% in the general population) and Russians underrepresented (16%in the biobank compared with the 25% in the general population).[73]

Another weakness is the limited depth of some subquestionnaires. For instance, a relatively brief food frequency questionnaire was administered with no detailed information on energy or nutrient intake; measurements of fasting glycaemic blood traits such as insulin level are available only for a limited set of samples. The limited depth of collected data can sometimes limit the number of projects in which the data can be used. However, more extensive questionnaires would have required even longer interview times and would have cost much more, possibly resulting in a smaller cohort size.

Can I get access to the data? Where can I find out more?

The EGCUT shares anonymous health, lifestyle, demographic and genetic data as well as biological materials (DNA, WBCs and plasma) for research and development projects. The data sharing is conducted in accordance with the regulations of the HGRA. A data application form can be found at www.biobank.ee.[4] The research project has to obtain approval from the Ethics Review Committee on Human Research of the University of Tartu[74] as well as approval from the EGCUT scientific committee.

The EGCUT has a policy for data sharing, which implies that for the EGCUT to share its resources, the applicant has to send the scientific results obtained from the research project conducted using the data or samples of the gene donors back to the Estonian Biobank.[3] These results will be added to the EGCUT database.

Future directions and development

The EGCUT has now established a compact biobank of considerable size. In the future, the emphasis will shift to studying the underlying mechanisms of common complex diseases. Another goal is informational enrichment through the additional collection of '-omics' data as well as phenotypic data. Concurrent with the data enrichment and research efforts, the EGCUT is working on building a system for implementing the information into the Estonian healthcare system. This is a complex process and requires multidisciplinary involvement from the medical, legal and research fields. As the first step, the genomes of 5000 gene donors will be sequenced in order to describe rare sequence variants and haplotypes among Estonians, and use them for the design of a new genotyping microarray, which can be added as supplemental markers in addition to common genome-wide and more rare exonic variants. Finally, this array will be used to genotype all gene donors in the Estonian Biobank and, when successful, everybody in Estonia aged between 35 and 65 years will be offered the test. With automated diagnosis support and risk estimation software incorporated into the eHealth system, we hope to provide disease risks and medication response information directly to the healthcare providers and thereby introduce genomics into healthcare in Estonia.

Supplementary data

Supplementary data are available at *IJE* online.

Funding

The establishment and maintenance of the Estonian Biobank has been supported by Targeted Financing from the Estonian Ministry of Science and Education [SF0180142s08], the Ministry of Social Affairs and the Ministry of Economic Affairs and Communications; the biobank and the Genome Center have been supported by the Development Fund of the University of Tartu (grant SP1GVARENG); the European Regional Development Fund has supported the Centre of Excellence in Genomics (EXCEGEN; grant 3.2.0304.11-0312) and through FP7 grants [278913, 306031, 313010, 245536, 269213, 212111, 201413].

Acknowledgements

The authors would also like to acknowledge all the recruiters and the participants of the EGCUT, the University of Tartu, the Ministry of Social Affairs, the Ministry of Science and Education, the Ministry of Economic Affairs and Communications, the Archimedes Foundation, the Estonian Biocentre, the Institute of Molecular and Cell Biology and the Centre for Ethics of the University of Tartu. The authors would also like to acknowledge the EGCUT technical personnel.

Other contributors: Annely Allik, Tarmo Annilo, Merli Hass, Atso-Heinar Jõks, Aidula-Taie Kaasik, Aime Keis, Erkki Leego, Merike Leego, Kadri Lilienthal, Kristjan Metsalu, Evelin Mihailov, Kairit Mikkel, Ene Mölder, Helja Niinemäe, Tiit Nikopensius, Mairo Puusepp, Steven Smit, Viljo Soo, Riin Tamm, Maris Teder-Laving and-Maris Väli-Täht.

Conflicts of interest: None declared.

References

1. Metspalu A. The Estonian Genome Project. *Drug Dev Res* 2004;**62**:97–101.
2. Metspalu A, Köhler F, Laschinski G, Ganten D, Roots I. [The Estonian Genome Project in the context of European genome research.] *Dtsch Med Wochenschr* 2004;**129**(**Suppl**):S25–28.
3. Human Genes Research Act. 2000. <http://www.geenivaramu.ee/for-scientists/human-genes-research-act.html> (7 January 2014, date last accessed).
4. Estonian Genome Center of the University of Tartu. 2013. <http://www.geenivaramu.ee/en/> (7 January 2014, date last accessed).
5. Population pyramid of Estonia: 2011. *Sta. Est Stat Bundesamt Deutschl*. 2011. <http://www.stat.ee/public/rahvastikupyramiid/> (7 January 2014, date last accessed).
6. Holle R, Happich M, Löwel H, Wichmann HE. KORA – a research platform for population based health research. *Gesundheitswesen* 2005;**67**(**Suppl. 1**):S19–25.
7. Health 2000. 2013. <http://www.terveys2000.fi/indexe.html> (4 June 2013, date last accessed).
8. Framingham. 2013. Original Cohort Framingham Heart Study. <http://www.framinghamheartstudy.org/about-fhs/background.php> (28 June 2013, date last accessed).
9. Estonian Health Insurance Fund. 2013. Family physicians. <http://www.haigekassa.ee/eng/service/physicians> (14 January 2014, date last accessed).
10. Knaapila A, Tuorila H, Silventoinen K *et al*. Food neophobia shows heritable variation in humans. *Physiol Behav* 2007;**91**:573–78.
11. Wichmann H-E, Kaaks R, Hoffmann W *et al*. The German National Cohort. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*. 2012;**55**:781-87.
12. Roenneberg T, Kuehnle T, Juda M *et al*. Epidemiology of the human circadian clock. *Sleep Med Rev* 2007; **11**:429–38.
13. Sheehan DV, Lecrubier Y, Sheehan KH *et al*. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998; **59**:22-33.
14. Gustavsson JP, Bergman H, Edman G *et al*. Swedish universities Scales of Personality (SSP): construction, internal consistency and normative data. *Acta Psychiatr Scand* 2000;**102**:217-25.
15. Costa PT, McCrae RR. Stability and change in personality assessment: the revised NEO Personality Inventory in the year 2000. *J Pers Assess* 1997;**68**:86-94.

16. ISO. ISO 9000 quality management <http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm> (20 June 2013, date last accessed).
17. World Health Organization. International Classification of Diseases (ICD). <http://www.who.int/classifications/icd/en/> (28 June 2013, date last accessed).
18. WHO Collaborating Centre. ATC/DDD Index. <http://www.whocc.no/atc_ddd_index/> (28 June 2013, date last accessed).
19. United Nations Educational, Scientific and Cultural Organization. International Standard Classification of Education. <http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx> (24 May 2013, date last accessed).
20. International Standard Classification of Occupations. <http://www.ilo.org/public/english/bureau/stat/isco/> (28 June 2013, date last accessed).
21. World Health Organization. Estonia. <http://www.who.int/countries/est/en/> (14 January 2014, date last accessed).
22. Estonian Ministry of the Interior. Population Register. <https://www.siseministeerium.ee/35796/> (13 January 2014, date last accessed).
23. National Institute for Health Development. Registers. <http://www.tai.ee/en/r-and-d/registers> (28 June 2013, date last accessed).
24. Estonian Health Insurance Fund. <http://www.haigekassa.ee/eng/> (28 June 2013, date last accessed).
25. eHealth. <http://www.e-tervis.ee/index.php/en/2012-07-22-13-35-31/organization> (17 September 2013, date last accessed)
26. Nelis M, Esko T, Magi R *et al*. Genetic structure of Europeans: a view from the North-East. *PLoS One* 2009;**4**:e5472.
27. Esko T, Mezzavilla M, Nelis M *et al*. Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Gene*. 2012;**20**:111–16.
28. BBMRI. <http://www.bbmri.eu/> (14 January 2014, date last accessed).
29. BBMRI-LPC. <http://www.bbmri-lpc.org/> (10 June 2013, date last accessed).
30. Home | Public Population Project in Genomics and Society. <http://p3g.org/> (14 January 2014, date last accessed).
31. Kettunen J, Tukiainen T, Sarin A-P *et al*. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* 2012; **44**:269–76.
32. Illig T, Gieger C, Zhai G *et al*. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 2010;**42**:137–41.
33. Codd V, Nelson CP, Albrecht E *et al*. Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* 2013; **45**:422–27.
34. Den Hoed M Eijgelsheim M, Esko T *et al*. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* 2013;**45**:621–31.
35. Berndt SI, Gustafsson S, Magi R *et al*. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* 2013;**45**:501–12.
36. Fernández-Rhodes L, Demerath EW, Cousimer DL *et al*. Association of adiposity genetic variants with menarche timing in 92,105 women of European descent. *Am J Epidemiol* 2013;**178**:451–60.
37. Gieger C, Radhakrishnan A, Cvejic A *et al*. New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011;**480**:201–08.
38. Okada Y, Sim X, Go MJ *et al*. Meta-analysis identifies multiple loci associated with kidney function-related traits in East Asian populations. *Nat Genet* 2012;**44**:904–09.
39. Pattaro C, Kottgen A, Teumer A *et al*. Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genet* 2012;**8**:e1002584.
40. Taal HR, St Pourcain B, Thiering E *et al*. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* 2013;**45**:713.
41. Bradfield JP, Taal HR, Timpson NJ *et al*. A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Gene*t* 2012*;***44**:526–31.
42. Codd V, Nelson CP, Albrecht E *et al*. Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* 2013;**45**:422–27.
43. Day-Williams AG, Southam L, Panoutsopoulou K *et al*. A variant in MCF2L is associated with osteoarthritis. *Am. J Hum Genet* 2011;**89**:446–50.
44. Deloukas, P, Kanoni S, Willenborg C *et al*. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2012;**45**:25–33.
45. Ellinghaus D, Ellinghaus E, Rajan P *et al*. Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci. *Am J. Hum Genet* 2012;**90**:636–47.
46. Jacquemont S, Reymond A, Zufferey F *et al*. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 2011;**478**:97–102.
47. Morris AP, Voight BF, Teslovich TM *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet;* 2012;**44**:981–90.
48. Nikopensius T, Annilo T, Jagomagi T *et al*. Non-syndromic tooth agenesis associated with a nonsense mutation in ectodysplasin-A (EDA). *J Dent Res* 2013;**92**:507–11.
49. Tsoi LC, Spain SL, Knight J *et al*. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Na. Genet* 2012*;***44**:1341–48.
50. Verhoeven VJM, Hysi PG, Wojciechowski R *et al*. Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet;*2013;**45**:712.
51. Võsa U, Vooder T, Kolde R *et al*. Meta-analysis of microRNA expression in lung cancer. *Int J Cancer* 2013;**132**:2884–93.
52. ArcOGEN collaborators. Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet* 2012;**380**:815–23.
53. Yang J, Loos RJ, Powell JE *et al*. FTO genotype is associated with phenotypic variability of body mass index. *Nature* 2012;**490**:267–72.
54. Walters RG, Coin L, Ruokonen A *et al*. Rare genomic structural variants in complex disease: lessons from the replication of associations with obesity. *PLoS One* 2013;**8**:e58048.
55. De Moor M, Costa P, Terracciano A *et al*. Meta-analysis of genome-wide association studies for personality. *Mol Psychiatry* 2012;**17**:337–49.

56. Terracciano A, Esko T, Sutan AR *et al*. Meta-analysis of genome-wide association studies identifies common variants in CTNNA2 associated with excitement-seeking. *Transl Psychiatry* 2011;**1**:e49.
57. Mõttus R, Real A, Allik J *et al*. Personality traits and eating habits in a large sample of Estonians. *Health Psychol* 2012;**31**:806–14.
58. Rietveld CA, Medland SE, Derringer J *et al*. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 2013;**340**:467–71.
59. Amin N, Hottenga J, Hansell N *et al*. Refining genome-wide linkage intervals using a meta-analysis of genome-wide association studies identifies loci influencing personality dimensions. *Eur J Hum Genet* 2013;**21**:876–82.
60. Middeldorp CM, de Moor M, McGrath L *et al*. The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data. *Transl. Psychiatry* 2011;**1**:e50.
61. Erhardt A, Akula N, Schumacher J *et al*. Replication and meta-analysis of TMEM132D gene variants in panic disorder. *Transl Psychiatry* 2012;**2**:e156.
62. Luciano M, Lopez L, de Moor M *et al*. Longevity candidate genes and their association with personality traits in the elderly. *Am J. Med Gene. B Neuropsychiatr Genet* 2011. doi:10.1002/ajmg.b.32013.
63. Ivanov M, Kals M, Kacevska M *et al*. In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic Acids Res* 2013;**41**:e72.
64. Lener MR, Gupta S, Scott R *et al*. Can selenium levels act as a marker of colorectal cancer risk? *BMC Cancer* 2013;**13**:214.
65. Kumar V, Westra H-J, Karjalainen J *et al*. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* 2013;**9**:e1003201.
66. Köttgen A, Albrecht E, Teumer A *et al*. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet* 2013;**45**:145–54.
67. Lokk K, Vooder T, Kolde R *et al*. Methylation markers of early-stage non-small cell lung cancer. *PLoS One* 2012;**7**:e39813.
68. Min JL, Nicholson G, Halgrimsdottir I *et al*. Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. *PLoS Genet* 2012;**8**:e1002505.
69. Scott RA, Lagon V, Welch RP *et al*. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 2012;**44**:991–1005.
70. Van der Harst P, Zhang W, Leach I *et al*. Seventy-five genetic loci influencing the human red blood cell. *Nature* 2012;**492**:369–75.
71. University of Tartu Estonian Gene Project. Gene Donor Consent Form. <http://www.geenivaramu.ee/for-donors/gene-donor-consent-form.html> (4 Jul 2013, date last accessed).
72. e-Estonia. X-Road. <http://e-estonia.com/components/x-road> (28 June 2013, date last accessed).
73. Statistics Estonia. <http://www.stat.ee/statistics> (10 June 2013, date last accessed).
74. Research Ethics Committee of the University of Tartu. <http://www.ut.ee/en/research/research-ethics-committee> (14 January 2014, date last accessed)