CORRESPONDENCE

Constructing a contemporary gene–environmental cohort: study design of the Yamagata Molecular Epidemiological Cohort Study

Journal of Human Genetics (2013) **58**, 54–56; doi:10.1038/jhg.2012.128; published online 22 November 2012

Development of common disorders involves a complex interplay of multiple factors. We may assume that many genes and many environmental factors are involved. Future gene–environmental cohort studies will no longer focus on single-gene abnormalities or major environmental risk factors, such as smoking in the case of lung cancer, which have been studied to date; rather, they will concentrate on the identification of risk factors that have not been discovered by conventional epidemiological methodologies.¹ Such risk factors preclude precise predictions because they will be revealed only after detailed analyses of the interactions among multiple factors—between genes and environmental factors, or among genes, for example. We may reasonably expect that combinations of these factors have an important role in the occurrence of a variety of common diseases, such as cancer, diabetes mellitus and cardiovascular diseases. These represent reasons why contemporary genomic cohort studies require the extensive collection of environmental and genetic information on a population level rather than from patients with specific disorders. Such a study design will enable researchers to formulate and test as many hypotheses as they can formulate. Launching a contemporary genomic cohort study requires enormous resources. Developed countries are investing in full-scale efforts to launch studies with cohorts comprising hundreds of thousands of local participants to meet the research objectives mentioned above. Such endeavors, such as the ongoing United Kingdom Biobank with 500 000 volunteers to date and the prospective cohort studies being planned in the United States, are termed ‘biobank’ projects.¹ In Japan, it is an urgent task to

build a contemporary gene–environmental cohort.

Yamagata University, with its long history of and experience with traditional cohort research projects that started some 30 years ago,^{2,3} started Yamagata Molecular Epidemiological Cohort Study, when its Global Center of Excellence (G-COE) program entitled ‘Formation of an International Network for Education and Research of Molecular Epidemiology’ (2008–2012) was approved by the Japanese Government. The summary of our study was shown Table 1. As of March 2012, 9000 participants (72.6% of the health check-up examinees; 3589 men and 5411 women; median age at baseline 64 years) donated their DNA to the study. Basically, the same study platform is shared with the Japan Multi-institutional Collaborative Cohort Study (J-MICC Study).⁴

We aim to conduct genome-wide gene–gene or gene–environmental interaction analysis using single-nucleotide polymorphisms (SNPs) in this study. It is an attractive way to identify genetic components that confer susceptibility to complex human diseases. However, individual hypothesis testing for SNP–SNP pairs, as in common genome-wide association studies (GWASs), involves difficulty in setting overall *P*-values due to the complicated correlation structure, namely, the multiple testing problem that causes unacceptable false-negative results.^{5,6} Specifically, there is difficulty in setting a genome-wide significance level using statistical methods such as Bonferroni correction, leading to prohibitively conservative results because they fail to successfully incorporate the correlation structure between each hypothesis. For instance, the total number of hypotheses for

gene–gene interaction is about 10^{11} – 10^{12} in standard GWAS data. Then the Bonferroni-corrected significance level must be considerably small; that is, the correction factor $<10^{-11}$. No efficient and universal multiple testing method to deal with such a huge set of hypotheses having a complicated correlation structure is proposed so far.⁷ The number of SNP–SNP pairs larger than the sample size, the so-called large *p* small *n* problem or the curse of dimensionality (that is, the number of predictors is larger than the sample size),^{8,9} precludes simultaneous analysis using multiple regression. To overcome these issues, our research team at Yamagata University developed an up-to-date method for ultrahigh-dimensional variable selection, termed sure independence screening (SIS), for appropriate handling of numerous SNP–SNP interactions by including them as predictor variables in logistic regression.⁷ This research team has also implemented the procedures in a software program, EPISIS, using the cost-effective GPGPU (general-purpose computing on graphics processing units) technology. The EPISIS program can complete an exhaustive search for SNP–SNP interactions in a standard GWAS data set within several hours. We plan to conduct SNP–SNP and SNP–environmental interaction analyses for associations of diseases or biological phenotypes or biomarkers using EPISIS software. Analysis using whole-genome information will be conducted in the future.

We obtained informed consent from each participant after a group orientation about the study, using participant-friendly materials that are available at our website (<http://gcoe.id.yamagata-u.ac.jp/jp/cohort/>). The main study protocol was approved by

Table 1 Summary of the study

Study periods	Baseline survey from January 2010 to March 2013 Second survey from April 2015 to March 2018 Participants would be followed until 2029
Target number of participants	20 000
Candidates ^a	Volunteers 40 to 74 years of age residing in the areas defined by municipal administrations and health check-up examinees visiting clinics run by municipalities in Yamagata Prefecture, Japan
Summary of surveys	<p>Lifestyle data^b</p> <p>Physical check-up and laboratory data</p> <p>Blood and urine samples</p> <p>Follow-up data^{d,e}</p> <p>The common questionnaire includes sleeping and exercise, alcohol drinking, smoking, psychological stress, medication and supplements, food intake frequency, family disease history, past disease history, and female reproductive history. In addition to the core questions, information on physical function, social support, and participation in community activities is included. This data are collected in the baseline survey and a secondary survey conducted ~5 years after the baseline survey. When a health check-up is conducted at the time of enrollment, the results of the health check-up^c are also subjected to analysis in this study. These data are collected in the baseline survey and the secondary survey.</p> <p>When the participants agree to donate blood for this study, 25 ml of blood are usually drawn into two 9-ml plain tubes for serum assays and a 7-ml EDTA-2Na-added tube for plasma and buffy coat assays. A urine sample is also collected. These samples are collected in the baseline survey and the secondary survey. End points are incident disease diagnosis, including cancers, cardiovascular diseases, strokes, and death from any cause. Biological phenotypes and biomarkers, including findings of laboratory tests, will also be used. Data are collected through the population-based cancer registry^f in Yamagata Prefecture, questionnaires mailed to the participants, questionnaires at repeated visits to health check-up facilities, notes from death certificates, information from health insurance data and second survey questionnaires.</p>
<i>Information explained to each participant</i> Purpose of this study Organization carrying out this study Study methods Privacy protection system Storage and analysis of survey materials and specimens Rights of study participants, benefits and disadvantages with this study Handling of analytical results, including genetic sequencing and method of disclosure of study results Handling of survey materials and specimens after the end of the study The fact that this is a joint study carried out together with the J-MICC study Access to study protocol Contact details for inquiries about this study	
<i>Consent to participate</i> Survey of health and lifestyle Donation of serum, plasma, DNA and urine Provision of test results of health checks and screenings Follow-up surveys Participation in other studies Long-term storage of specimens and participation in genetic sequencing	

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; GTP, guanosine-5'-triphosphate; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

^aMost of the health check-up examinees are National Health Insurance (NHI) beneficiaries. In Japan, NHI is used by farmers, self-employed individuals, pensioners and their dependents.

^bThe core questions in the questionnaire are common to the J-MICC Study.

^cHealth Check-up Items to be Collected includes date of health check-up, physical condition (height, weight and blood pressure), urinalysis (protein and glucose), serum tests (HDL cholesterol, LDL cholesterol, triacylglycerol, ALT, AST and γ GTP), blood cell counts (red blood cells, hemoglobin and hematocrit) and plasma tests (blood glucose and hemoglobin A1c).

^dThe incidence data from the mail questionnaires, questionnaires at repeated visits, death certificates, health insurance data and second survey questionnaires are confirmed by the hospital record or the disease registries. Data from the disease registries will be collected annually after the data are available.

^eIf the participants move their resident record outside Yamagata Prefecture, their data collection ends as of the time when they move out.

^fThe cancer registry in Yamagata Prefecture is of sufficient quality; in 2008, the rate of death certificate notification was 18.5% and the rate of death certificate only was 5.9%. The data from the registry are included in 'Cancer Incidence in Five Continents (IARC Scientific Publications)'. Yamagata prefecture also founded population-based registries for cardiovascular diseases and cerebrovascular diseases in 2010 and these registries will be also used.

the Ethics Committee at Yamagata University School of Medicine.

We described the study design of the gene-environmental cohort study to unveil how genetic and environmental factors jointly influence the risk of common human disease development. The strength of our study is as follows: the capacities of constructing a large cohort study, such as the size of 20 000; the study platform with high commonality to other study groups; the novel useful method, statistical method and the software; and the high-quality cancer registration. A study cohort of 20 000 participants is being gathered for contemporary genomic analysis during the G-COE program. After the G-COE program, the size of cohort is planned to expand to cover all of Yamagata prefecture. The ultimate target is a study cohort that is comparable in size to the United Kingdom and United States genomic cohorts. For this purpose, we are collaborating with the large cohort studies in Japan, such as the J-MICC study, as well as expanding our cohort size. We also plan to share the

information related to our newly developed genomic analysis⁷ with other study groups.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid from the Global Center of Excellence program of the Japan Society for the Promotion of Science. The contributors of the study until March in 2012 are shown in Appendix.

Yamagata University Genomic Cohort Consortium (YUGCC) and Hiroto Narimatsu

Advanced Molecular Epidemiology Research Institute, Faculty of Medicine, Yamagata University, 2-2-2 Iida-nishi, Yamagata, Japan
E-mail: hiroto-narimatsu@umin.org

- 1 Manolio, T. A., Bailey-Wilson, J. E. & Collins, F.S. Genes, environment and the value of prospective cohort studies. *Nat. Rev. Genet.* **7**, 812–820 (2006).

- 2 Konta, T., Hao, Z., Abiko, H., Ishikawa, M., Takahashi, T., Ikeda, A. *et al.* Prevalence and risk factor analysis of microalbuminuria in Japanese general population: the Takahata study. *Kidney Int.* **70**, 751–756 (2006).
- 3 Tominaga, M., Eguchi, H., Manaka, H., Igarashi, K., Kato, T. & Sekikawa, A. Impaired glucose tolerance is a risk factor for cardiovascular disease, but not impaired fasting glucose. The Funagata Diabetes Study. *Diabetes Care* **22**, 920–924 (1999).
- 4 Hamajima, N. The Japan Multi-Institutional Collaborative Cohort Study (J-MICC Study) to detect gene-environment interactions for cancer. *Asian Pac. J. Cancer Prev.* **8**, 317–323 (2007).
- 5 Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413–417 (2005).
- 6 Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10**, 392–404 (2009).
- 7 Ueki, M. & Tamiya, G. Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. *BMC Bioinformatics* **13**, 72 (2012).
- 8 Johnstone, I. M. & Titterton, D. M. Statistical challenges of high-dimensional data. *Philosophical transactions. Ser. A Math. Phys. Eng. Sci.* **367**, 4237–4253 (2009).
- 9 Catchpole, D. R., Kennedy, P., Skillicorn, D. B. & Simoff, S. The curse of dimensionality: a blessing to personalized medicine. *J. Clin. Oncol.* **28**, e723–e724, author reply e725 (2010).

APPENDIX

This study was conducted under the auspices of the following contributors at Yamagata University, Yamagata, Japan

Yamagata University Global COE Program Leader: Takamasa Kayama

Lead Principal Investigator: Akira Fukao

Steering Committee: Hidetoshi Yamashita (Dean, Faculty of Medicine), Isao Kubota (Respiratory and Cardiovascular Diseases Research Center), Takeo Kato (Metabolic and Degenerative Diseases Research Center), Chifumi Kitanaka (Oncology Research Center), Shinya Sato (Respiratory and Cardiovascular Diseases Research Center), Yoshiyuki Ueno (Oncology Research Center).

Cohort Study: Hiroto Narimatsu, Kyoko Shibata, Akiko Miura, Rina Inoue, Ai Numazawa, Kahori Kudo, Yoko Aita, Noriko Umezawa, Yuko Saito, Yumi Takahashi and Yuka Suzuki (Cohort Management Unit), Katsumi Otani, Atsushi Hozawa and Li Shao (Department of Public Health), Masatsugu Orui (Yamagata Prefectural Tsuruoka Hospital), Kei Honma (Department of Ophthalmology and Visual Science), Atsuko Kobayashi, Yuka Kanoya, Takiko Hosoya, Ikuko Suzuki, Mariko Otake, Yuko Morikagi, Akiko Sekimata, Manami Hiraka, Yumi Matsuda, Chika Sato, Yoko Takeda, Yoko Matsunami, Tatsuya Horie, Shiho Sato, Mizue Inoue and Kaoru Baba (School of Nursing), Tsuneo Konta, Yoko Shibata and Tetsu Watanabe (Department of

Cardiology, Pulmonology and Nephrology), Takafumi Saito, Naohiko Makino (Department of Gastroenterology), Makoto Daimon, Toru Kawanami, Manabu Wada, Toshihide Oizumi, Chifumi Iseki and Yoshimi Takahashi (Department of Neurology, Hematology, Metabolism, Endocrinology and Diabetology).

Data Analysis: Gen Tamiya, Masao Ueki, Tomohiro Nakamura, Jamiyansuren Jambaldorj and Satoko Araki (Genomic Information Analysis Unit).

Data Management: Kazuei Takahashi and Kazuo Goto (Data Management Unit).

Specimen Management: Osamu Nakajima (Specimen Management Unit).

Patent Management and Commercialization: Kimishige Ishizaka (COME Center Inc.).