Omics research project on prospective cohort studies from the Tohoku Medical Megabank Project

Seizo Koshiba1,2 | Ikuko Motoike1,3 | Daisuke Saigusa1,2 | Jin Inoue1,2 |
Matsuyuki Shirot1,2 | Yasutake Katoh1 | Fumiki Katsuoka1,2 | Inaho Danjoh1,2 |
Atsushi Hozawa1,2 | Shinichi Kuriyama1,2,4 | Naoko Minegishi1,2 | Masao Nagasaki1,2,3 |
Takako Takai-Igarashi1,2 | Soichi Ogishima1,2 | Nobuo Fuse1,2 | Shigeo Kure1,2 |
Gen Tamiya1,2,5 | Osamu Tanabe1,2 | Jun Yasuda1,2 | Kengo Kinoshita1,3,6 |
Masayuki Yamamoto1,2

1Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan
2Graduate School of Medicine, Tohoku University, Sendai, Japan
3Graduate School of Information Sciences, Tohoku University, Sendai, Japan
4International Research Institute of Disaster Science, Tohoku University, Sendai, Japan
5RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
6Institute of Development, Aging and Cancer, Tohoku University, Sendai, Japan

Correspondence
Masayuki Yamamoto and Seizo Koshiba, Tohoku Medical Megabank Organization, Tohoku University, Japan.
Emails: masiyamamoto@med.tohoku.ac.jp; koshiba@megabank.tohoku.ac.jp

Funding information
Ministry of Education, Culture, Sports, Science and Technology; Japan Agency for Medical Research and Development, Grant/Award Number: JP15km0105001, JP16km0105001, JP17km0105001, JP15km0105002, JP16km0105002 and JP17km0105002; Platform Program for Promotion of Genome Medicine, Grant/Award Number: JP15km0405001, JP16km0405001 and JP17km0405001; Advanced Genome Research and Bioinformatics Study to Facilitate Medical Innovation, Grant/Award Number: JP16km0405203 and JP17km0405203; Japan Science and Technology Agency

Population-based prospective cohort studies are indispensable for modern medical research as they provide important knowledge on the influences of many kinds of genetic and environmental factors on the cause of disease. Although traditional cohort studies are mainly conducted using questionnaires and physical examinations, modern cohort studies incorporate omics and genomic approaches to obtain comprehensive physical information, including genetic information. Here, we report the design and midterm results of multi-omics analysis on population-based prospective cohort studies from the Tohoku Medical Megabank (TMM) Project. We have incorporated genomic and metabolomic studies in the TMM cohort study as both metabolome and genome analyses are suitable for high-throughput analysis of large-scale cohort samples. Moreover, an association study between the metabolome and genome show that metabolites are an important intermediate phenotype connecting genetic and lifestyle factors to physical and pathologic phenotypes. We apply our metabolome and genome analyses to large-scale cohort samples in the following studies.
Tohoku Medical Megabank (TMM) Project has been founded to operate prospective cohort studies in Japan’s Miyagi and Iwate prefectures. Both regions were affected by the Great East Japan Earthquake on 11 March 2011 (Kuriyama et al., 2016), and the cohort studies of TMM are aimed at supporting community medicine in these earthquake-damaged regions by establishing next-generation medical systems, such as personalized health care and personalized medicine. For this purpose, the Tohoku Medical Megabank Organization (ToMMo) at Tohoku University and Iwate Tohoku Medical Megabank Organization (IMM) at Iwate Medical University have been cooperatively conducting prospective cohort studies and constructing an integrated biobank.

Tohoku Medical Megabank started two types of cohort studies in 2013: a population-based adult cohort study named “the TMM Community-Based Cohort Study (TMM CommCohort Study)” and a birth and three-generation cohort study named “the TMM Birth and Three-Generation Cohort Study (TMM BirThree Cohort Study)” (Kuriyama et al., 2016). At the first stage (baseline) of these cohort studies (from 2013 to 2017), more than 150,000 participants have been successfully recruited. In 2017, we started the second stage of our cohort studies called “TMM Repeat Assessment Center-based Survey during the Second Period.”

One of the characteristics of our cohort studies is the implementation of a wide variety of omics and genomic analyses within a conventional questionnaire-based cohort study; molecular profiling of each participant is important for investigating the associations of genetic and environmental factors with phenotypes. Another reason for the incorporation of omics and genomic analyses is to avoid rapid depletion of the TMM biobank samples. In the TMM biobank based on these two cohort studies, an analytical center within the biobank conducts standard omics and genomic analyses and provides/shares data with the science community. TMM biobank provides samples upon request that are not covered by these standard analyses. This biobank operation has been referred to as the integrated biobank (Kuriyama et al., 2016).

In terms of omics and genomic analyses, genome research has already been recognized as a major method of cohort studies (Cupples et al., 2007; Leitsalu et al., 2015; Scholtens et al., 2015; Sudlow et al., 2015; Wichmann, Gieger, & Illig, 2005). In fact, we have already analyzed the whole-genome sequences of more than 2,000 participants with next-generation sequencing systems (Nagasaki et al., 2015; Yamaguchi-Kabata et al., 2017). Single nucleotide polymorphism (SNP) data of more than 10,000 participants have also been analyzed by our custom array system (Japonica Array®) (Kawai et al., 2015). In addition, metabolome analysis is becoming an indispensable method to investigate molecular phenotypes in cohort studies (Gieger et al., 2008; Harada et al., 2016; Kettunen et al., 2012; Loeffler et al., 2015; Rhee et al., 2013; Saw et al., 2017; Schmidt et al., 2015; Shin et al., 2014; Tigchelaar et al., 2015). The metabolome is an intermediate phenotype connecting genotypes and phenotypes and mediates the effects of environmental and lifestyle factors to phenotypes. Therefore, metabolome analysis is a complementary and suitable method for investigating the effects of these factors on phenotypes. In particular, the application of metabolome analysis to samples from families participating in our three-generation cohort study is expected to be important for the estimation of gene–environmental influences.

In the last decade, population-based cohort studies start collecting genome and omics information for all or a subset of participants. For example, one of the major cohort studies, LifeLine Deep study in Netherlands, has obtained methylome and transcriptome data from more than 1,000 participants, in addition to the genome and metabolome information (Tigchelaar et al., 2015). Recently, such cohort studies aiming at molecular phenotypes are operated not only in Europe and America but also in the other regions of the world such as Asia (Saw et al., 2017). As genetic factors as well as environmental factors such as diets and lifestyles are different among various regions in the world, molecular-based phenotypes in each region are important for elucidating the contributions of each genetic or environmental factor to the health of populations. Especially, Japanese is known for longevity and information of their genetic and environmental factors should be important for the international molecular epidemiology. Our TMM cohort project is the first and largest population-based prospective cohort study including genome and omics investigation in Japan. TMM cohort project also performs a three-generation cohort study targeting children. This family-based cohort information including fetal environments is very effective for investigating interactions between genetic and environmental factors and their effects on common diseases.

Although omics and genomic analyses play important roles in recent cohort studies, some problems remain in implementing these analyses to population-based cohort studies (Elliott, Peakman, & Biobank, 2008; Hirayama et al., 2015). One of the major problems is the difficulty of sample collection for omics analyses during health surveys in cohort studies. Although individual genome datum does not change during a lifetime, profiles of other omics targets, such as metabolites, are influenced by many factors, such as diet, disease, aging and genetic polymorphism. Hence, detailed information on these factors must be collected to evaluate their effects on omics data. In contrast, the quality of omics data, except for genome data, largely depends on the quality of collected samples because targets of omics analyses, such as metabolites, are generally not stable compared to...
the stability of the genome. Therefore, suitable protocols for sample collection, especially blood plasma collection, must be developed and evaluated under various limitations that are characteristic of cohort studies before starting large-scale studies.

In this report, we describe the study design of the ToMMo omics analyses, focusing mainly on metabolome analysis in our cohort studies. One of the characteristics of our approach is that we performed the nontarget type of metabolome analysis of plasma using two analytical methods: nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). This approach is suitable for achieving our goals to perform high-throughput analysis of the metabolic profiles of cohort participants and to elucidate biomarkers for personalized healthcare. We also conducted the nontarget type of proteome analysis by MS to elucidate blood protein profiles. For these omics analyses, we developed and evaluated blood sample collection protocols in our cohort studies. We report the results of our omics research in the first stage of our cohort studies and will discuss the future research plan.

2 COHORT DESIGN AND SUBJECTS FOR OMICS STUDY

TMM has been conducting two prospective cohort studies: a population-based adult cohort study and a birth and three-generation cohort study (Kuriyama et al., 2016). The TMM CommCohort Study targets adult individuals (aged 20 and older) and has recruited a total of 84,073 participants from 2013 to 2016. In contrast, the TMM BirThree Cohort Study targets pregnant women and their families, including fathers, grandparents and other family members of the fetuses. This study has recruited a total of 73,001 participants, including 32,086 children, from 2013 to 2017. On 1 June 2017, we started the TMM Repeat Assessment Center-based Survey during the Second Period and are planning to recruit participants from the baseline surveys for follow-up measurements of health data. Details of this repeat assessment study will be described elsewhere.

The ToMMo omics study basically targets adult participants of the TMM baseline cohort studies (Figure 1). Although we did not select underage participants for the current omics study, we are planning to investigate them when the follow-up

**FIGURE 1** Overview of the ToMMo omics research project in the Tohoku Medical Megabank (TMM) project
survey proceeds. The cohort studies of the TMM project are approved by the ethics committees of Tohoku University and Iwate Medical University. The ToMMo omics study is approved by the ethics committee of Tohoku University. All adult participants have signed an informed consent.

3 | SAMPLE COLLECTION AND STORAGE FOR OMICS ANALYSES

Before initiating our cohort studies, we investigated protocols for the collection, transfer, processing and storage of blood samples suitable for omics analyses. In particular, we carefully considered sample collection and transfer processes because the quality of blood samples is largely influenced by time and temperature during these processes. Therefore, we investigated the influences of these factors on the quality of blood samples using several samples collected from volunteers and created protocols for the sample collection and transfer processes (Saigusa et al., 2016). We also considered which type of blood collection method should be used. Based on these analyses, we concluded that plasma is better for omics analyses than serum because several proteins and metabolites are not stable during the clotting process of serum. Nonetheless, we also collected serum samples in our cohort studies because serum is a standard sample for some blood test items and some laboratories still use serum as a sample for biological analyses. Through these analyses, we evaluated the protocol for cohort sample collection for omics analyses as follows (Figure 2).

In the case of plasma, blood samples are collected using Venoject II® tubes containing EDTA-2Na (Terumo Corporation). After collection, sample tubes are immediately inverted 10 times and are stored at 4°C. These sample tubes are transported to our biobank laboratory using refrigerated containers with temperature data loggers. Total transport time is within 8 hr for most samples collected in the Miyagi prefecture. The transported tubes are centrifuged at 2,330 × g for 10 min at 4°C. The plasma fraction is transferred to the liquid handling machine (Freedom EVO®, Tecan) and dispensed to MATRIX® 1.0-ml 2D barcoded screw tubes (Thermo Scientific). The number of the dispensed tubes is basically four per blood sample, and the plasma volume in each tube is approximately 700 μl. These samples are administered by TMM biobank ID and stored at −80°C.

In the case of serum, blood samples are collected using the Venoject II® tubes with clot activator (Terumo Corporation). After clotting at room temperature, sample tubes are stored at 4°C and are transported to the laboratory in a similar manner to plasma transport. The transported tubes are centrifuged at 1,700 × g for 10 min at 4°C, and the serum fraction is transferred to 2D barcoded screw tubes (700 μl/tube, 4 tubes/participant). These samples are also administered by TMM biobank ID and stored at −80°C.

The sample preparation process in the ToMMo biobank laboratory is operated by the laboratory information management system (LIMS) using a barcode system and TMM biobank ID. The storage process of these samples is managed using an automated storage platform and biobank data management system (Brooks Life Science System). For omics analyses, the plasma sample in each dispensed tube (700 μl) is further divided into six tubes (approximately 120 μl per tube), which are managed using ToMMo omics ID and a barcode system.

4 | PROTEOME ANALYSIS

For the proteome analysis of plasma samples, we applied nontarget type MS analysis because information related to the blood protein profiling of cohort participants is an important basis for a wide variety of studies, such as biomarker discovery and multi-omics analysis. To this end, we investigated the protocols of sample preparation and MS measurements to establish the method for nontargeted proteome analysis (Figure 3a). As a result, we adopted the method in which plasma samples are first denatured, reduced and alkylated. The resulting samples are then digested in tandem by lysylendopeptidase and trypsin. After desalting, the samples are analyzed using liquid chromatography followed by tandem mass spectrometry (LC-MS/MS) systems (Thermo Fisher Scientific). Applying this protocol, we measured a total of 501 plasma samples with triplicate measuring. Peptide
identification from the obtained data was performed using SequestHT and Mascot search engines with the UniProt human proteome set (April 2014) as the reference for protein sequences. The identified peptide data were integrated using Proteome Discoverer software (version 1.4: Thermo Fisher Scientific).

5 | METABOLOME ANALYSIS

The two major aims of our metabolome study using plasma samples of cohort participants can be summarized as follows. One aim is to comprehensively and quantitatively elucidate the standard profile of Japanese blood metabolites, whereas the other is to investigate biomarkers for diseases and personal health conditions by means of a large-scale association study of metabolites with other omics/genomics and cohort data. Hence, we have applied the nontarget type method to analyze plasma metabolites. In our analyses, the metabolites in plasma were examined by a combination of NMR spectroscopy and MS.

5.1 | NMR metabolome analysis

Detailed protocols for the NMR metabolome analysis of plasma have been described (Koshiba et al., 2016). In brief, metabolites were extracted from 200 μl of plasma and suspended in a sodium phosphate buffer. All NMR experiments were performed at 298 K (25°C) on a Bruker 600 MHz spectrometer (Figure 3b). Standard 1D NOESY and CPMG spectra were obtained from each plasma sample (Beckonert et al., 2007). The samples were analyzed using the Chenomx NMR Suite 8.0 (Chenomx), and metabolites were manually quantified using the target profiling approach in the feasibility study as described below. However, we developed an in-house automatic quantification program to analyze a number of cohort samples and migrated to the new analytical protocol (Aoki et al., in preparation). Whereas this protocol currently quantifies 37 metabolites, we are now developing our metabolite database to expand the quantification target of this program. The statistical profile of our metabolome data has been released as a public database called “Japanese Multi Omics Reference Panel (jMorp)” as will be described in the following section.

5.2 | MS metabolome analysis

We have also developed a protocol for the nontarget analysis of plasma metabolites in cohort participants by MS spectrometry (Figure 3b) (Saigusa et al., 2016). Metabolites were extracted from a 50-μl plasma sample. We developed an automatic sample extraction system based on a Microlab® STARlet robot system (Hamilton, Reno, NV, USA) to prepare MS samples. This system can automatically prepare 192 samples within 2 hr. MS analysis was performed by exploiting two types of liquid chromatography–mass spectrometry (LC-MS) systems. For positive ion mode, we used a UHPLC-QTOF/MS system (an Acquity Ultra Performance LC I-class and a Synapt G2-Si QTOF MS with electrospray ionization [Waters]). A C18 column (Acquity HSS T3; Waters) was used for LC separation. The other system is an LC-FTMS system for negative ion mode, consisting of a NANOSPACE SI-II HPLC (Shiseido, Tokyo) and a Q Exactive Orbitrap MS system (Thermo Fisher Scientific, San Jose, CA, USA). LC separation was performed using a HILIC column (ZIC-pHILIC; Sequant, Darmstadt). The total measurement time was 15 min for one sample.

5.3 | Quality control of samples by metabolome analyses

Although we have collected and analyzed cohort blood samples based on our established protocols as described above,
it seems inevitable that the quality of several samples will be lowered by some accidents during cohort studies. Hence, we have excluded data derived from these low-quality samples from our omics database based on the results of our metabolome analyses (Saigusa et al., 2016).

In practice, we have examined the concentrations of some metabolites (particularly glucose and lactate) of each plasma sample and excluded the data if the quality of samples was insufficient. We first check samples whose glucose concentrations from NMR data were lower than those from blood test (if ratio is less than 0.7) and whose lactate concentrations are higher than the average (more than 2 SD). Second, we investigate a history of each outlier and, if we identify causes of outliers, we exclude such samples from our database. For example, we have excluded samples whose plasma samples were stored more than 1 day after the blood collection due to accidents. We are also developing quality control markers for serum samples.

6 | FEASIBILITY STUDY FOR TOMMO OMICS ANALYSES

Before executing large-scale omics analyses of cohort samples, we conducted a feasibility study using approximately 500 plasma samples collected in the TMM CommCohort Study (Koshiba et al., 2016; Saigusa et al., 2016). For the multi-omics analysis, we selected samples whose whole-genome sequences were already available in ToMMo (Nagasaki et al., 2015). We performed metabolome and proteome analyses based on our developed protocols.

In the case of MS metabolome analyses, we successfully measured plasma samples using a nontarget LC-MS system. We detected more than 1,000 peaks of features in total and identified 250 metabolites. We also measured metabolites in plasma samples by NMR and identified and quantified 37 metabolites. The quantification quality of NMR is higher than that of MS analysis, as the quantified plasma glucose values by NMR correlate well with those of laboratory blood tests than those from LC-MS do (Figure S1). Although fewer metabolites are detected by NMR than by LC-MS/MS, NMR is useful for elucidating the standard profile of blood metabolites. Based on these results, we conclude that metabolome analysis is suitable and adequate for high-throughput analysis of large-scale cohort samples. Because metabolome analyses by NMR and LC-MS are complementary, we decided to conduct regular metabolome analysis of large-scale cohort samples using both NMR and MS systems.

In contrast, we performed nontarget proteome analysis on 501 plasma samples. We detected 51,916 peptide signals and identified more than 7,873 protein types in total. On average, 284 protein types were detected for one sample by our proteome analysis protocol. To the best of our knowledge, this work represents the first example that has obtained three layers of omics data, that is, metabolome, proteome and whole-genome data, of samples collected from a population-based cohort study.

In this regard, we found two problems related to nontarget MS proteome analysis of plasma samples from large-scale cohort studies. One issue is the long measurement time or low throughput. The measurement time of proteome analysis is approximately 3 hr for one sample, which is much longer than that of metabolome analysis (approximately 10 min [NMR] to 15 min [MS] per sample). The other problem is the reproducibility of measurements. We measured the same sample a total of three times for all cohort plasma samples and found that on average, only 45.8% of the detected proteins were observed all three times. Our results indicate that nontarget proteome analysis of plasma samples is not suitable for analyzing many plasma samples from large-scale cohort studies. We envisage that proteome analysis is more suitable for analyzing a limited number of samples to investigate specific phenotypes. We are planning to develop a more reliable method for proteome analysis and will apply the protocol for large-scale cohort samples in the near future.

In conclusion, we performed a feasibility study for omics analyses in large-scale cohort studies. We found that metabolome analysis is suitable for high-throughput analysis of plasma samples. The measurement time of metabolome analyses is short enough for a large-scale cohort study, and the reproducibility and quality of the data are good enough to analyze several tens of thousands of plasma samples collected from large-scale cohort studies. In contrast, proteome analysis is difficult currently in large-scale cohort samples and seems to be more suitable for investigating selected samples to verify the results of association studies of other omics (genome, transcriptome, metabolome, etc.) and cohort data.

7 | JAPANESE MULTI-OMICS REFERENCE PANEL (JMORP)

We have developed a new database referred to as the “Japanese Multi Omics Reference Panel” (jMorp: http://jmor.megabank.tohoku.ac.jp) so that the information obtained from omics analyses of our cohort samples is accessible worldwide. This database provides a standard profile of blood metabolites and proteins in a Japanese population (Tadaka et al., 2017).

The jMorp database consists of concentration distributions of 37 metabolites identified by NMR, distributions of peak intensities of 257 metabolites by LC-MS, and observed frequencies of 256 proteins (Table 1). The first version of jMorp was released in 2015, and jMorp has been updated in 2016 and 2017. The latest version includes the analysis of
5,093 plasma samples (male: 2,077, female: 3,016) collected from participants of our cohort studies. The main interface of jMorp is the top search page, which consists of a search window and a table that lists the identified metabolites and proteins (Figure 4a). Users can search for molecules of interest using a text-based search function, which accepts molecular names, ToMMo molecular IDs, or mass-to-charge ratios \((m/z)\).

For each metabolite, we have provided three types of information: distribution of concentrations obtained from NMR data (distribution of peak intensities for MS data), age dependency of distribution and a list of correlated metabolites, which are graphically viewed by an interactive network viewer (Figure 4b,c). These three kinds of information are classified by sex (all, male or female) to provide information on the sex dependency of each metabolite distribution and correlation with other metabolites. In addition, we also included the results of our first metabolome–genome association study (MGWAS) for five metabolites (Koshiba et al., 2016).

We have also provided proteome analysis results. Two types of information are provided for each protein: the

| Data               | Method   | Source | Number of samples | Number of molecular items |
|--------------------|----------|--------|-------------------|--------------------------|
| Basic information  | –        | –      | 5,093             | –                        |
| Metabolome         | NMR      | Plasma | 5,093             | 37                       |
| Metabolome         | LC-MS    | Plasma | 1,312             | 257                      |
| Proteome           | LC-MS    | Plasma | 501               | 256                      |

BMI, body mass index; LC-MS, liquid chromatography followed by mass spectrometry; NMR, nuclear magnetic resonance.

**FIGURE 4** Interface of jMorp. (a) Top search page. Note that users can select a specific molecule by a text search function or by selecting the molecule from a list of molecules. (b) Example of metabolite data (phenylalanine). Left graph is the distribution of the metabolite concentration, and right graph is the age-dependent distribution of the metabolite. (c) Example of a network viewer for the correlation network among metabolites.
detection rate of each protein in the target participants and peptide sequences detected by MS. The detected peptide information contains not only peptides derived from reference alleles of the genome but also those derived from alternative alleles containing nonsynonymous genetic variants.

Links to other public databases are also included for each entry in jMorp: Human Metabolome Database, Kyoto Encyclopedia of Gene and Genomes, and LIPID MAPS for metabolites, and UniProt for proteins. These database sites provide a wide variety of information, helping users understand each molecule.

Our jMorp database is expected to be used for the many applications, including (1) as a healthy reference in many research fields, such as case–control studies, (2) as a reference for the detection of compounds or medicines not found in healthy people and (3) for the investigation of new biological processes based on metabolite networks. We will regularly update the contents of the jMorp database in accordance with the progress of our omics studies. We will also include information from our association studies, such as associations with genetic variations, questionnaires from the cohort or blood test values.

Although we provide the statistical information of metabolome and proteome in jMorp database, the multi-omics data for each individual can also be shared upon request. Researchers can request distributions of these individual data to the office of Materials and Information Management in ToMMo (http://www.dist.megabank.tohoku.ac.jp/). The requests are examined and judged by Sample and Data Access Committee for the distribution of TMM samples and information.

While ToMMo conducts three-layered multi-omics analyses (genome, proteome and metabolome) of plasma samples, IMM investigates the epigenome analysis of blood cells and recently reported another three-layered multi-omics analysis (genome, epigenome and transcriptome) for purified blood cells (monocytes, CD4+ T cells and neutrophils) from more than 100 cohort participants (Hachiya et al., 2017). IMM has released these results on another public multi-omics database, the integrative DNA methylation database (iMethyl; http://imethyl.iwate-megabank.org/). ToMMo and IMM are now developing protocols for more than three-layered multi-omics analyses of cohort samples.

8 | ASSOCIATION STUDY

In the TMM prospective cohort studies, we also conducted many assessments of baseline (first) cohort studies: questionnaires (including those on disaster experiences), collection of biospecimens (blood, urine, saliva, etc.), detailed blood tests, physical and medical examinations and magnetic resonance imaging (Kuriyama et al., 2016). From 2017 onward, follow-up reassessment (second) cohort studies are conducted and similar assessments are performed to the first cohort study. Because one of the main purposes of our omics study is to identify effective biomarkers for personal healthcare, we are investigating the associations of our metabolome and proteome data with a wide variety of cohort assessment data as their data cleaning processes are now underway.

ToMMo has already released the first results of the association study between the plasma NMR metabolome data and the whole-genome sequence data obtained from 512 participants (Koshiba et al., 2016). We identified five significant associations of metabolites with nonsynonymous variants, four of which were known to exist in enzymes involved in metabolic disorders (Figure 5). An intriguing observation derived from the structural analyses is that these nonsynonymous

FIGURE 5 Association study of metabolome and genome data. Manhattan plots for metabolic traits. Each metabolite and its associated gene are indicated. The line indicates a suggestive genome-wide significance level with a p-value of $7.08 \times 10^{-9}$. 
variants are located in peripheral regions of the catalytic sites or related regulatory domains, indicating that these variants retain only a moderate impact on their enzymatic activities but reside in the population with a certain frequency or are stably inherited in this population. However, as shown in Figure 6, two individuals with larger changes in the levels of the metabolite phenylalanine have also been identified; these individuals retained rare variants in the phenylalanine hydroxylase gene that cause nonsynonymous variations located near the catalytic sites. This observation indicates that these variations cause substantial changes in metabolite levels (Koshiba et al., 2016). These results demonstrate that variant frequency, structural location and phenotypic effect correlate with each other in human populations.

This association study implies that metabolome analysis should play an important role in future cohort and medical studies (Figure 7). In general, variants that moderately affect enzymatic activity seem to be nearly neutral in their effects on individual fitness, resulting in their solid accumulation during evolution. These genetic variations (SNPs) eliciting moderate effects are identified in our association study of genetics and cohort omics studies. We surmise that these variations are involved in a variety of diseases, suggesting that the accumulation of moderate variants significantly affects metabolic individuality and susceptibility to lifestyle-related diseases.

It has been shown that even moderately affecting variants can be associated with a significant change in metabolite levels, indicating that the metabolome is an ideal intermediate phenotype between genome and clinical phenotypes (Adamski & Suhre, 2013). Of note, relationships between metabolites and disease phenotypes have been investigated extensively by a wide variety of clinical and basic medical studies (DeBerardinis & Thompson, 2012; Duarte, Diaz, & Gil, 2014; Friedrich, 2012; Griffin & Shockcor, 2004; Lindon & Nicholson, 2014; Nicholson et al., 2012; Quinones & Kaddurah-Daouk, 2009); the mechanisms underlying these metabolite–disease relationships can be elucidated by multiple biological approaches. In contrast, direct associations of such moderate genetic variants with disease phenotypes are usually difficult as the...

**FIGURE 6** Relationship among variant, allele frequency and structure. (a) Scatter plot of the concentration distribution of plasma phenylalanine. Heterozygotes of rs118092776, rs79931499 and rs746203167 are shown in orange, red and pink, respectively. Others are shown in gray. (b) Mapping of the three nonsynonymous variants on the structure of rat PAH (phenylalanine hydroxylase). Three residues corresponding to the position of each nonsynonymous variant are shown by a sphere model.
associations are too weak to be detected, resulting in the need for huge numbers of samples and costs for identification. We further speculate that multifactorial diseases may be provoked by a combination of these genetic changes or SNPs that cause moderate molecular phenotypes, each of which has a limited contribution to the ultimate phenotype.

In conclusion, MGWAS allows us to quantitatively and qualitatively identify these molecular phenotypes and enables the detection of associations between metabolite changes and genetic variants even with a limited number of samples. Thus, MGWAS analysis is an attractive approach that links molecule phenotypes to the pathogenesis of multifactorial diseases. Furthermore, we would like to emphasize that metabolome research is a powerful approach that connects genome cohort studies or epidemiology to experimental sciences, such as biochemistry, biophysics and molecular biology, which aim to show the mechanistic insights of gene–disease relationships.

9 | FUTURE PLAN

Based on the omics studies linked to the first stage of the TMM cohort projects, we are now planning a second-stage omics study in collaboration with the follow-up cohort examination. Our main omics approach will be metabolome analysis in association with genome and transcriptome analyses. In addition, our collaborator, IMM, will conduct epigenome analysis.

The second stage of metabolome analysis will improve the quality of the metabolic profiles of cohort participants and is expected to elucidate biomarkers that are useful for personal healthcare. In particular, we are planning to use the metabolic profiles to classify cohort participants because such stratification will be important and valuable to establish segments that are effective in the evaluation of potential disease risk. To this end, we are planning to increase the number of people and metabolites examined in the study.

Whereas the main target of our omics study in the first stage was adult participants of the TMM CommCohort Study, in the second stage, we will also target family participants of the TMM BirThree Cohort Study. The TMM cohort project offers a great advantage in investigating gene–environmental interactions because more than 10,000 trio families were recruited in the TMM BirThree Cohort Study. These families included approximately 200 complete sets of three-generation families (i.e., four grandparents, two parents and a child). We are hoping to elucidate the contribution of each genetic and environmental factor to metabolic profiles through analyses.

We also have started metabolome analysis of the samples collected in the “TMM Repeat Assessment Center-based Survey” to investigate the metabolic profile changes of each participant relative to those in the first survey. We are also planning to conduct metabolome analyses of mononuclear cells and immortalized lymphocytes of the participants, which will be useful to investigate the effects of genetic variants, drugs or a wide variety of stresses (Hayton, Maker, Mullaney, & Trengove, 2017; Niu & Wang, 2015). Furthermore, we will continue to develop protocols for proteome analysis using these mononuclear cells and immortalized lymphocytes. In addition, we are now developing protocols for the target metabolome analysis of cohort samples with a gas chromatography-MS system. Finally, we also have started transcriptome analysis of blood samples not only to assess the functional contributions of genetic variants found in introns.
and regulatory regions of genes but also to identify biomarkers for diseases (Battle et al., 2014; Cummings et al., 2017; Kayano et al., 2016; Satoh, Kino, & Niida, 2015).

10 CONCLUSION

In the TMM cohort study, we have been conducting omics analysis and have been developing protocols for metabolome and proteome analyses that are suitable for high-throughput plasma sample analysis of large-scale cohorts. We have generated standard profiles of blood metabolites and proteins in a Japanese population and have released them as a public database, jMorp. We have also examined the association between plasma metabolites and SNPs, which proves that metabolome analysis is a powerful approach for investigating the association between genetic variations and phenotypes. In the second stage of the ToMMo project, we will continue metabolome analysis to delineate the profiles of blood metabolites in most participants and the associations of genetic and environmental factors in pathophysiological conditions.

ACKNOWLEDGMENTS

We thank all the volunteers who participated in this study. We thank the members of ToMMo at Tohoku University for contributing to the establishment of the genome cohort and biobank and for helping with omics analyses (http://www.megabank.tohoku.ac.jp/english/a171201). This work was supported in part by the Tohoku Medical Megabank Project from MEXT, Japan Agency for Medical Research and Development (AMED; Grant Numbers JP15km0105001, JP16km0105001, JP17km0105001, JP15km0105002, JP16km0105002 and JP17km0105002), Platform Program for Promotion of Genome Medicine (P3GM; Grant Numbers JP15km0405001, JP16km0405001 and JP17km0405001), Advanced Genome Research and Bioinformatics Study to Facilitate Medical Innovation (Grant Numbers, JP16km0405203 and JP17km0405203), Center of Innovation Program from Japan Science and Technology Agency (JST), Project for Promoting Public Utilization of Advanced Research Infrastructure (MEXT) and Practical Research Project for Life-Style related Diseases including Cardiovascular Diseases and Diabetes Mellitus (AMED).

CONFLICT OF INTEREST

The authors declare no competing financial interests.

ORCID

Seizo Koshiba http://orcid.org/0000-0003-2001-1949

REFERENCES

Adamski, J., & Suhr, K. (2013). Metabolomics platforms for genome wide association studies—Linking the genome to the metabolome. Current Opinion in Biotechnology, 24, 39–47. https://doi.org/10.1016/j.copbio.2012.10.003

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., … Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research, 24, 14–24. https://doi.org/10.1101/gr.155192.113

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature Protocols, 2, 2692–2703. https://doi.org/10.1038/nprot.2007.376

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., … MacArthur, D. G. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science Translational Medicine, 9, pii: eaai5209. https://doi.org/10.1126/scitranslmed.aai5209

Cupples, L. A., Arruda, H. T., Benjamin, E. J., D’Agostino, Sr, R. B., Demissie, S., DeStefano, A. L., … Atwood, L. D. (2007). The Framingham Heart Study 100K SNP genome-wide association study resource: Overview of 17 phenotype working group reports. BMC Medical Genetics, 8 (Suppl 1), S1. https://doi.org/10.1186/1471-2350-8-S1-S1

DeBerardinis, R. J., & Thompson, C. B. (2012). Cellular metabolism and disease: What do metabolic outliers teach us? Cell, 148, 1132–1144. https://doi.org/10.1016/j.cell.2012.02.032

Duarte, I. F., Díaz, S. O., & Gil, A. M. (2014). NMR metabolomics of human blood and urine in disease research. Journal of Pharmaceutical and Biomedical Analysis, 93, 17–26. https://doi.org/10.1016/j.jpba.2013.09.025

Elliott, P., Peakman, T. C., & Biobank, U. K. (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. International Journal of Epidemiology, 37, 234–244. https://doi.org/10.1093/ije/dym276

Friedrich, N. (2012). Metabolomics in diabetes research. Journal of Endocrinology, 215, 29–42. https://doi.org/10.1530/JOE-12-0120

Gieger, C., Geistlinger, L., Altmaier, E., Hrabe de Angelis, M., Kronenberg, F., Meitinger, T., … Suhr, K. (2008). Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. PLoS Genetics, 4, e1000282. https://doi.org/10.1371/journal.pgen.1000282

Griffin, J. L., & Shockcor, J. P. (2004). Metabolic profiles of cancer cells. Nature Reviews Cancer, 4, 551–561. https://doi.org/10.1038/nrc1390

Hachiya, T., Furukawa, R., Shiwa, Y., Ohmomo, H., Ono, K., Katsuoka, F., … Shimizu, A. (2017). Genome-wide identification of inter-individually variable DNA methylation sites improves the efficacy of epigenetic association studies. npj Genomic Medicine, 2, 11. https://doi.org/10.1038/s41525-017-0016-5

Harada, S., Takebayashi, T., Kurihara, A., Akiyama, M., Suzuki, A., Hatakeyama, Y., … Tomita, M. (2016). Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. Environmental Health and Preventive Medicine, 21, 18–26. https://doi.org/10.1007/s12199-015-0494-y

Hayton, S., Maker, G. L., Mullaney, I., & Trengove, R. D. (2017). Experimental design and reporting standards for metabolomics studies of mammalian cell lines. Cellular and Molecular Life Sciences, 74, 4421–4441. https://doi.org/10.1007/s00018-017-2582-1
Hirayama, A., Sugimoto, M., Suzuki, A., Hatakeyama, Y., Enomoto, A., Harada, S., … Takebayashi, T. (2015). Effects of processing and storage conditions on charged metabolomic profiles in blood. *Electrophoresis*, https://doi.org/10.1002/elps.201400600

Kawai, Y., Mimori, T., Kojima, K., Nartai, N., Danjoh, I., Saito, R., … Nagasaki, M. (2015). Japonica array: Improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *Journal of Human Genetics*, 60, 581–587. https://doi.org/10.1038/jhg.2015.68

Kayano, M., Higaki, S., Satoh, J. I., Matsumoto, K., Matabara, E., Takikawa, O., & Niida, S. (2016). Plasma microRNA biomarker detection for mild cognitive impairment using differential correlation analysis. *Biomarker Research*, 4, 22. https://doi.org/10.1186/s40364-016-0076-1

Kettunen, J., Tukiainen, T., Sarin, A. P., Ortega-Alonso, A., Tikkanen, E., Lytyikäinen, L. P., … Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44, 269–276. https://doi.org/10.1038/ng.1073

Koshiya, S., Motoike, I., Kojima, K., Hasegawa, T., Shirota, M., Saito, T., … Yamamoto, M. (2016). The structural origin of metabolic quantitative diversity. *Scientific Reports*, 6, 31463. https://doi.org/10.1038/srep31463

Kuriyama, S., Yagi, N., Nagami, F., Arai, T., Kawaguchi, Y., Osumi, N., … Yamamoto, M. (2016). The Tohoku Medical Megabank Project: Design and Mission. *Journal of Epidemiology*, 26, 493–511. https://doi.org/10.2188/jeaJE20150268

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M. L., Alavere, H., Snieder, H., … Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, 44, 1137–1147. https://doi.org/10.1093/ije/dyv268

Lindon, J. C., & Nicholson, J. K. (2014). The emergent role of metabolic phenotyping in dynamic patient stratification. *Expert Opinion on Drug Metabolism & Toxicology*, 10, 915–919. https://doi.org/10.1517/17425255.2014.922954

Loffler, M., Engel, C., Ahnert, P., Allermann, D., Arelin, K., Baber, R., … Thiery, J. (2015). The LIFE-Adult-Study: Objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC Public Health*, 15, 691. https://doi.org/10.1186/s12889-015-1983-z

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., … Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, 6, 8018. https://doi.org/10.1038/ncomms9018

Nicholson, J. K., Holmes, E., Kinross, J. M., Darzi, A. W., Takats, Z., & Lindon, J. C. (2012). Metabolic phenotyping in clinical and surgical environments. *Nature*, 491, 384–392. https://doi.org/10.1038/nature11708

Niu, N., & Wang, L. (2015). In vitro human cell line models to predict clinical response to anticancer drugs. *Pharmacogenomics*, 16, 273–285. https://doi.org/10.2217/pgs.14.170

Quinones, M. P., & Kaddurah-Daouk, R. (2009). Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. *Neurobiology of Diseases*, 35, 165–176. https://doi.org/10.1016/j.nbd.2009.02.019

Rhee, E. P., Ho, J. E., Chen, M. H., Shen, D., Cheng, S., Larson, M. G., … Gerszten, R. E. (2013). A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metabolism*, 18, 130–143. https://doi.org/10.1016/j.cmet.2013.06.013

Saigusa, D., Okamura, Y., Motoike, I. N., Katoh, Y., Kurosawa, Y., Saijyo, R., … Yamamoto, M. (2016). Establishment of protocols for global metabolomics by LC-MS for biomarker discovery. *PLoS ONE*, 11, e0160555. https://doi.org/10.1371/journal.pone.0160555

Satoh, J., Kino, Y., & Niida, S. (2015). MicroRNA-Seq data analysis pipeline to identify blood biomarkers for Alzheimer’s disease from public data. *Biomark Insights*, 10, 21–31. https://doi.org/10.4137/BMLS25132

Saw, W. Y., Tansoto, E., Begum, H., Zhou, L., Zou, R., He, C., … Teo, Y. Y. (2017). Establishing multiple omics baselines for three Southeast Asian populations in the Singapore Integrative Omics Study. *Nature Communications*, 8, 653. https://doi.org/10.1038/s41467-017-00413-x

Schmidt, J. A., Rinaldi, S., Ferrari, P., Carayol, M., Achaïnthe, D., Scalbert, A., … Travis, R. C. (2015). Metabolic profiles of male meat eaters, fish eaters, vegetarians, and vegans from the EPIC-Oxford cohort. *American Journal of Clinical Nutrition*, 102, 1518–1526. https://doi.org/10.3945/ajcn.115.111989

Scholtens, S., Smidt, N., Swertz, M. A., Bakker, S. J., Dotinga, A., Vonk, J. M., … Stolk, R. P. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology*, 44, 1172–1180. https://doi.org/10.1093/ije/dyu229

Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsieck, J., Santos, R., Huang, J., … Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46, 543–550. https://doi.org/10.1038/ng.2982

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases and old age. *PLoS Medicine*, 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Tadaka, S., Saigusa, D., Motoike, I. N., Inoue, J., Aoki, Y., Shirota, M., … Kinoshita, K. (2017). jMorp: Japanese Multi Omics Reference Panel. *Nucleic Acids Research*, 46, D551–D557. https://doi.org/10.1093/nar/gkx978

Tsigchelea, E. F., Zhernakova, A., Dekens, J. A., Hermes, G., Baranska, A., Mujagic, Z., … Feskens, E. J. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *British Medical Journal Open*, 5, e006772. https://doi.org/10.1136/bmjopen-2014-006772

Wichmann, H. E., Gieger, C., & Illig, T.; MONICA/KORA Study Group (2005). KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67 (Suppl 1), S26–S30. https://doi.org/10.1055/s-2005-858226

Yamaguchi-Kabata, Y., Yasuda, J., Tanabe, O., Suzuki, Y., Kawame, H., Fuse, N., … Yamamoto, M. (2017). Evaluation of reported pathogenic variants and their frequencies in a Japanese population based on a whole-genome reference panel of 2049 individuals. *Journal of Human Genetics*, 63, 213–230. https://doi.org/10.1038/sj hg.2017.128

**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information section at the end of the article.

**How to cite this article:** Koshiba S, Motoike I, Saigusa D, et al. Omics research project on prospective cohort studies from the Tohoku Medical Megabank Project. *Genes Cells*. 2018;23:406–417. https://doi.org/10.1111/gtc.12588