

## Review Article

# Establishment of Integrated Biobank for Precision Medicine and Personalized Healthcare: The Tohoku Medical Megabank Project

Nobuo Fuse<sup>1)</sup>, Mika Sakurai-Yageta<sup>2)</sup>, Fumiki Katsuoka<sup>1)</sup>, Inaho Danjoh<sup>1)</sup>, Ritsuko Shimizu<sup>1),3)</sup>, Gen Tamiya<sup>1),4)</sup>, Fuji Nagami<sup>5)</sup>, Hiroshi Kawame<sup>2)</sup>, Shinichi Higuchi<sup>5)</sup>, Kengo Kinoshita<sup>1),6)</sup>, Shigeo Kure<sup>1),7)</sup>, and Masayuki Yamamoto<sup>1),8)</sup>

**Abstract:**

The Tohoku Medical Megabank (TMM) project was established to provide creative reconstruction of the Tohoku area that suffered from a huge earthquake and ensuing tsunami (the Great East Japan Earthquake, GEJE). TMM aims to establish two large-scale genome cohorts and an integrated biobank managing biospecimen and related information. It supports community medicine by establishing next-generation medical systems through a combination of the prospective genome cohort studies with a total of 150,000 participants and genomic medicine. The strategies for genome analyses in TMM are to develop an elaborate genome reference panel by means of high-fidelity Japanese whole-genome sequence, to design custom single nucleotide polymorphism (SNP) arrays based on the reference panel, and to obtain genotype data for all the TMM cohort participants subsequently. Disease-associated genomic information and omics data, including metabolomics and microbiome analysis, provide an essential platform for precision medicine and personalized healthcare (PHC). Ethical, legal, and social issues (ELSI) and education are important for implementing genomic medicine. The major considerations of ELSI regarding each participant of the cohort studies are the respect for the autonomy and the protection of privacies. Moreover, developing and provide human resources not only for the TMM project but also for the social implementation of precision medicine and PHC is required. We started a pilot study of the return of genomic results for familial hypercholesterolemia (FH) as a target disease. TMM aims to establish solid platforms that support precision medicine and PHC based on the genomic and omics information and environmental and lifestyle factors of the individuals, which is one of the most advanced medical care beyond the evidenced-based medicine in the near future.

**Key Words:**

genomic medicine, personalized healthcare, precision medicine, Tohoku Medical Megabank Project, whole-genome reference panel, Japonica array, clinical sequencing

## Introduction

A huge earthquake (the Great East Japan Earthquake, GEJE) with tremendous tsunami waves damaged wide areas of the northeastern coast of Japan on March 11, 2011 <sup>(1)</sup>. Bad circumstances, imbalanced nutrition <sup>(2)</sup>, and the lack of daily activities influenced on people who suffered from the earthquake <sup>(3)</sup>. To revive medical services in those damaged regions, we proposed the Tohoku Medical Megabank (TMM) Project to establish next-generation medical systems through a combination of the prospective genome cohort studies with 150,000

participants and medical genetics in Miyagi and Iwate Prefectures. To conduct the TMM Project in the two prefectures, the Tohoku Medical Megabank Organization (ToMMo) was established at Tohoku University in Miyagi Prefecture, and the Iwate Tohoku Medical Megabank Organization (IMM) was established at Iwate Medical University in Iwate Prefecture. We constructed three platforms: *i*) biobank, *ii*) two types of prospective genome cohorts, and *iii*) genome and omics analysis (**Figure 1**). The purpose of our project is to develop the fundamental infrastructure consisting of genome and cohort information for the establishment of precision medicine

<sup>1)</sup>Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. <sup>2)</sup>Department of Education and Training, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. <sup>3)</sup>Department of Molecular Hematology, Graduate School of Medicine, Tohoku University, Sendai, Japan. <sup>4)</sup>Statistical Genetics Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. <sup>5)</sup>Department of Public Relations and Planning, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. <sup>6)</sup>Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Japan. <sup>7)</sup>Department of Pediatrics, Graduate School of Medicine, Tohoku University, Sendai, Japan. <sup>8)</sup>Department of Medical Biochemistry, Graduate School of Medicine, Tohoku University, Sendai, Japan

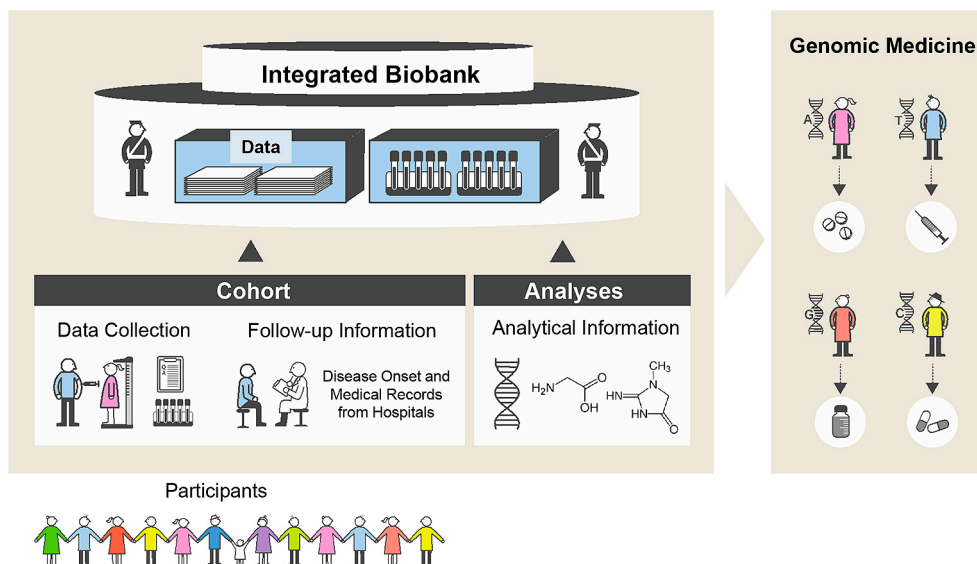
Corresponding author: Masayuki Yamamoto, masiyamamoto@med.tohoku.ac.jp

JMA J. 2019;2(2):113-122

Received: March 3, 2019 / Accepted: April 12, 2019 / Advance Publication: August 6, 2019 / Published: September 4, 2019

Copyright © Japan Medical Association

## Integrated Biobank to Genomic Medicine



**Figure 1.** TMM establishes “integrated biobank” consisting of biospecimens, information of cohort studies, and genome-omics analyses. The integrated biobank is an important research infrastructure for implementing genomic medicine.

and personalized healthcare (PHC) in Japan.

### Establishment of Two TMM Cohorts

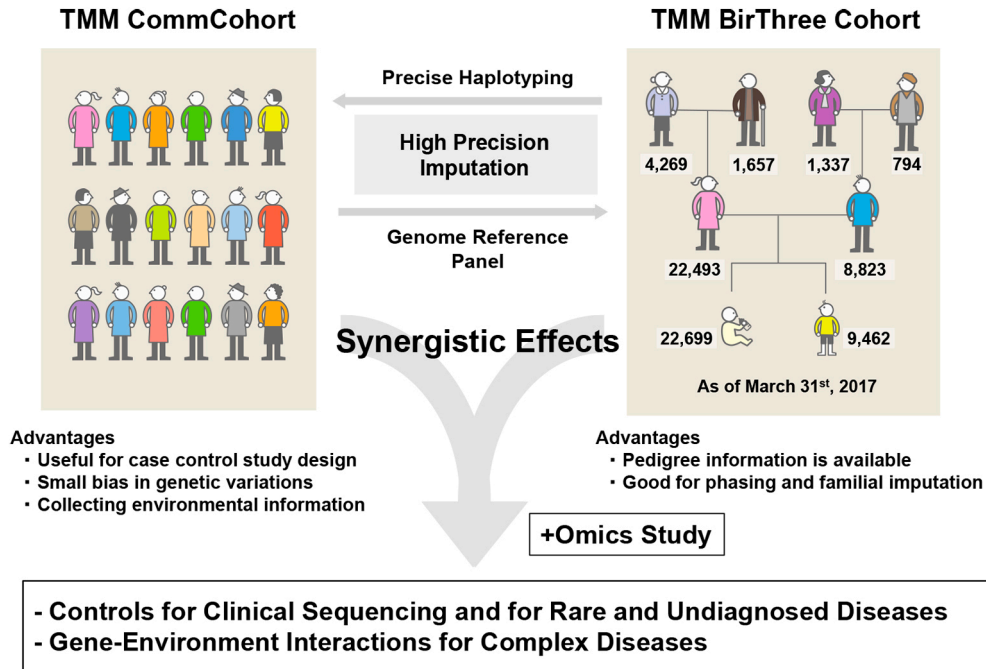
TMM has been working to establish two cohort studies <sup>(4)</sup> (Figure 2). One is the TMM community-based cohort study (TMM CommCohort Study), in which a total 84,073 participants (52,212 and 31,861 from Miyagi and Iwate, respectively), 20 years or older (mean age, 61 years), were recruited by March 2016. The other is the TMM Birth and Three-Generation Cohort Study (TMM BirThree Cohort Study), which included pregnant mothers, their husbands (or partners), their parents, the newborn babies, and their elder siblings. In the TMM BirThree Cohort Study, a total of 73,500 participants were recruited from the obstetrics hospitals and clinics in the Miyagi Prefecture by July 2017 from 22,493 families. Through the TMM BirThree Cohort Study, we are hoping to collect longitudinal data from the pregnancy and postnatal periods to adulthood.

To conduct a smooth recruitment of residents for the TMM CommCohort Study and TMM BirThree Cohort Study, we established seven local community support centers in Miyagi Prefecture and four satellites in Iwate Prefecture. We performed a number of examinations, such as central aortic blood pressure, body composition, bone density, grip strength, leg strength, oral health, respiratory function, hearing, and eye examinations and carotid ultrasound imaging. Moreover, in the TMM project, an “integrated biobank” con-

sisting of biospecimens, health and clinical information, and genome-omics data for the research infrastructure was established.

PHC is the concept in which individual health should be managed by means of the individual’s specific characteristics. PHC actively utilizes the individual’s genomic and omics profiles by exploiting appropriate methods and interventions to prevent various acquired diseases. In addition, genomic medicine is an emerging medical discipline that utilizes genome information of individuals for clinical care. The use of genomic medicine is now expanding to the prediction of the safety and effectiveness of medical treatments and risk assessments of each individual. Precision medicine and PHC are becoming two of the most advanced but feasible forms of medical care with multiple approaches to personal care. Precision medicine and PHC have the potential to detect diseases at earlier stage and as such easily make the treatment effective. Full implementation of precision medicine and PHC should involve risk assessment, prevention, detection, diagnosis, treatment, and management of diseases. Because TMM project is based on two prospective genome cohort studies, we are focusing especially on risk assessment and prevention.

Theoretically, a prospective cohort study treats all diseases that occur in the participants, who are ostensibly healthy people. However, our major focus is on the common diseases that involve an interaction between genetic and environmental factors. Therefore, we informed the cohort participants of the major diseases we are especially focusing on: cancers, cardio-



**Figure 2.** TMM establishes two cohorts, TMM CommCohort and TMM BirThree Cohort. Taking advantages of these cohorts and their synergistic effects, we provide the reference panel as control data for clinical diagnosis and establish risk prediction scores of complex diseases by gene-environment interaction studies.

vascular disorders, diabetes, mental diseases (PTSD and depression), dementia, and respiratory diseases, such as asthma and chronic obstructive pulmonary disease. In the TMM BirThree Cohort, the target diseases also include atopic dermatitis, autism, and low birth weight of babies and preeclampsia of mothers <sup>(4)</sup>.

To firmly establish PHC, the association between disease phenotype, environmental factors, and genotype-omics data must be deeply investigated. In the case of Mendelian diseases, highly penetrant disease-responsible variants can be easily identified. In contrast, in the case of common and multifactorial diseases, a large number of common and rare variants associated with such diseases must be sought for the risk estimation. A number of single nucleotide polymorphisms (SNPs), which are associated with many human traits or diseases, were already reported. However, these disease-associated SNPs cannot sufficiently explain the heritability of disease susceptibility. The discrepancies between observed heritability and the contributions to disease susceptibility by the disease-associated SNPs have been referred to as “missing heritability” <sup>(5)</sup>.

Nowadays, the advent of next-generation sequencing and SNP array analysis with genotype imputation enabled us to acquire information of all variants from a population. In this regard, TMM has been focusing on providing the essential infrastructure, which is a large set of genome variation data of the general residents. This genome reference panel would contribute

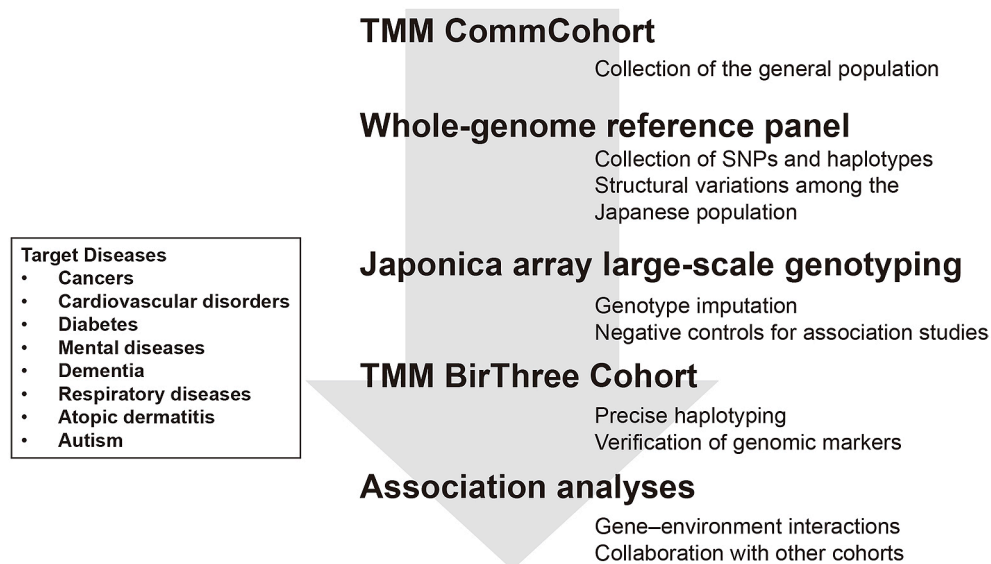
as control data for collaborative association studies. Thus, TMM provides an analytical platform that medical doctors and researchers are able to perform collaborative studies by using genome and medical data under secure protection of privacy and personal information.

The project protocols were reviewed and approved by the Ethics Committee of Tohoku University Graduate School of Medicine and by the Ethics Committee of Iwate Medical University, and written informed consent was obtained from all of the participants.

### Strategy of Genome and Omics Analyses on Genome Cohorts

We constructed two kinds of cohorts, i.e., TMM CommCohort and TMM BirThree Cohort. Our strategy for the genome and omics analyses intends to maximize the benefits of these two different cohorts, which recruited healthy volunteers. We planned to construct reference panels of Japanese genome and omics first mainly using the TMM CommCohort Study. Based on the information derived from the reference panels, we planned to prepare Japonica array for large-scale genotyping of the Japanese population and, subsequently, to utilize the TMM BirThree Cohort for verification of the results of genome analyses conducted using the Japonica array. The TMM BirThree Cohort may also be useful for the con-

## Project Strategy for Genome Analyses



**Figure 3.** Our strategy of genome analyses and target diseases. A whole-genome reference panel is developed by using TMM CommCohort and applied for genotype imputation subsequent to large-scale genotyping by Japonica array<sup>TM</sup>. The TMM BirThree Cohort is utilized for the collection of accurate haplotype information.

struction of precise haplotype collections capitalizing their familial information. Thus, while these two cohorts are managed independently, the combination of the two types of cohorts in the genome and omics analyses will be effective in revealing important mechanistic insights of complex diseases caused by gene-environment interactions (**Figure 3**).

Based on the above considerations, we first selected approximately 1,000 participants from the TMM CommCohort and performed whole-genome sequencing (WGS) analysis. To clarify the genetic variations in the Japanese population, we constructed a Japanese whole-genome reference panel based on the WGS data<sup>(6)</sup>. As will be described precisely below, we then developed a custom SNP array specially designed for the Japanese population to expedite collection of genotype data<sup>(7)</sup>. Indeed, we are expediting the collection of genomic information of 150,000 participants by using the SNP array, as genome data of the TMM BirThree Cohort will be significant to improve accuracy of haplotypes. It should be noted that accurate haplotype collection could be used for precise genotype imputations with the SNP array.

As for the omics analyses, we conducted transcriptome, epigenome, proteome, metabolome, and metagenome analyses. By using these omics and other data, we are trying to clarify the gene-environment interactions within the Japanese population and accumulate solid lines of evidence for the realization of future medicine in Japan<sup>(8)</sup>. These fundamental strategies of genome-omics analyses share certain similarity

with those of deCODE study in Iceland, as well as other modern genome cohort studies, such as UK Biobank and Finland Biobank. Meanwhile, it should be noted that the TMM project is unique as it is operating two closely related cohorts in an elaborate combination and also conducting the birth and three-generation cohort successfully, even though similar birth and three-generation cohort attempts in the United States and the United Kingdom were cancelled as a result of the shortage of participants<sup>(9)</sup>. Since the inception of the TMM project in 2012, we have been steadily analyzing the participants' genome and omics following the strategy summarized here. For more details on the genome-omics analyses of our project, please see Yasuda et al.<sup>(10)</sup> and Koshiba et al.<sup>(11),(12)</sup>.

### Genotyping Using Japonica Array<sup>TM</sup>

To establish the precision medicine, it is mandatory to collect precise genome data from all cohort participants. In this regard, although WGS gives rise to detailed information of each genome, its use for the entire cohort samples is not realistic, as it is too expensive. To solve this contradiction, a microarray approach coupled with a genotype imputation method is often employed in various genome cohorts worldwide<sup>(13)</sup>. Polymorphic genome reference panels with haplotype phasing are necessary for the genotype imputation of microarray data. As patterns of linkage disequilibrium (LD) block are distinct

**Table 1.** Contents of Japonica Array™ v2.

Category	Number of SNPs
Tag SNPs (including X chromosome)	632,186
Pharmacogenomic markers (ADME)	2,020
Y chromosome	606
Mitochondria	104
NHGRI GWAS catalog	11,171
HLA	6,914
Untagged functional SNPs	3,990
Total	659,328

among populations<sup>(14), (15)</sup>, imputation accuracy heavily depends on the quality of both reference panel and microarray data. To obtain highly accurate performance of the imputation, the genome reference panel constructed with genome data from thousands of individuals and genotyping array designed to include polymorphic markers reflecting the LD block pattern are necessary.

We started the development of Japanese-specific full custom array, referred to as Japonica Array™ (JPA) in 2013<sup>(7)</sup>. Until the end of 2018, we produced three full custom arrays, JPAv1, JPAv2, and JPAv3. Currently, JPAv2 is mainly used for the data acquisition of TMM cohort samples (**Table 1**). JPAv1 and JPAv2 have already been placed on the market by Toshiba. JPAv3 is focused on disease-related SNPs identified in the Japanese population as well as tag SNPs.

We collected SNP data from our cohort participants by using JPAv1 and JPAv2 that are optimized for genotyping of the Japanese population. Genotyping data, in 2018 fiscal year, were obtained from >68,000 participants. Mainly using JPAv2, we will complete the collection of microarray genotype data from 114,000 participants by the end of March 2019.

We set up analytical pipelines consisting of wet laboratory- and informatics-based processes in ToMMo, and most of the genotyping assays were performed in the ToMMo facility. Our wet laboratory equips Laboratory Information Management System (LIMS) for quality control and has the capacity to acquire JPA data for more than 6,000 samples per month. The acquired data is automatically transferred into supercomputer and imputed by using the 3.5KJPNv2 reference panel subsequent to multiple steps of genotyping and quality control analyses. At present, imputed dataset consisting of individual-level genotype data is released on approximately 32,000 TMM participants, including genomic data obtained by SNP arrays other than JPA.

We are planning to complete genotyping of all participants (more than 150,000) by March 2021 and to release full dataset after imputation in line with our data-sharing policy. In addition, we are planning to share our analytical pipelines for our collaborators who have various case samples. These ap-

proaches would promote the study of gene-environment interaction and the development of disease risk scores, bringing about the realization of genomic medicine.

## Risk Analysis Based on Genome Information

A true set of genetic and environmental factors conferring the risk of various complex diseases is valuable for the prediction of disease onset or prognosis. However, disease risk predictions are still in their infancy, chiefly due to uncaptured genetic components, biased environmental measures, complex non-additive combinatorial effects, insufficient experimental/statistical methods, and so on<sup>(16)</sup>. Large-scale prospective genomic cohort with biobank, such as TMM, will provide an opportunity to overcome such limitations in the disease risk predictions<sup>(9)</sup>. From identical populations in TMM, one can obtain a more comprehensive assessment of the full range of human genetic variation and less-biased environmental measurements<sup>(17)</sup>. Unlike scores based on any single risk factor, some sophisticated multivariate models including ours<sup>(18), (19)</sup> are now beginning to yield more accurate scores than before even under the large-p small-n conditions in standard genome-wide association studies (GWASs).

One of our methods, Iwate polygenic model (iPGM), efficiently utilizes GWAS data along with environmental factors using a framework of linear mixed model<sup>(19)</sup>. Another method that we developed, smooth-threshold multivariate genetic prediction (STMGP), is a statistical machine-learning method based on an analogous framework of ridge regression that gets around  $p \gg n$  problem from yielding the best prediction performance (or area under the curve, AUC) among known methods in the other public cohorts<sup>(18)</sup>.

Using these approaches as core algorithms coupled with a large-scale genetic and environmental data from TMM, we will carry out our disease risk predictions in Miyagi and Iwate. We will then expand the risk prediction to the other areas of Japan. Our effort could overcome the missing heritability problem to realize disease risk predictions by incorporating a

comprehensive set of disease-contributing factors: additive, nonadditive genetic factors, environmental factors, and their interactions. Showing very good agreement with our approach toward these improvements in data collection and risk analyses, polygenic risk scores (PRSs) generated by the DNA microarray analyses and other clinical examinations have been highly expected to predict risks of multifactorial diseases. We theorize that the time to exploit genome analysis data through PRSs tests and to predict disease risks for the complex diseases in clinic should come soon <sup>(20)</sup>.

## Metagenomic Analysis

We are continuously exposed to environmental microbial species as well as environmental chemicals, pollutions, and ultraviolet. Metagenomics is a direct genetic analysis of the genomes of microorganisms that reside in an environmental sample. Recent advances in high-throughput genome-sequencing technique and in bioinformatics allow us to verify precise taxonomic classification in the microbiome communities. Experimental lines of evidence were gathered to show that the patterns of human microbiomes are different between healthy people and patients with specific diseases. Thus, characterizations of the human microbiomes, especially symbiotic microbiomes residing in our body parts, and analyses of the roles that they play in health and disease conditions are emerging as a growing scientific interest.

We developed an oral metagenomics as one of the pipelines. Characterizations of multiple microbiomes are expected to contribute to health and disease monitoring. Multiple microorganisms of different species inhabit the plaque in the oral cavity. Buildup of plaque is an essential characteristic of oral infectious diseases <sup>(21)</sup>, such as periodontal disease and dental caries, which affect nearly all ages and all ethnicity of people worldwide. Certain species of oral microbiomes are thought to play important roles in modulating host response not only in the oral cavity but also in the whole body, by acting as pathogens and/or changing metabolic pathways <sup>(22), (23)</sup>. Comprehensive analysis by metagenomics together with host genomics and metabolomics is an important strategy for public health surveillance and risk prediction.

We collected plaque and saliva from more than 25,000 participants and analyzed their microbial profiles using the 16S ribosomal RNA gene high-throughput sequencing. We are planning to investigate not only the associations of oral microbial profiles with host phenotypes but also the cross relationships of multiple phenotypes accumulating in the TMM project. We theorize that this comprehensive approach brings about the understanding as to what is more influential to the onset of diseases among environmental, lifestyle, and genetic factors.

## Ethical, Legal, and Social Issues (ELSI)

The TMM project is a huge, complex, and multidirectional project. At the planning stage of the project, the adequacy of the genome cohort project as a reconstruction program for the damage caused by the GEJE was widely discussed. The frameworks of the project, especially those of the cohort studies, were arranged to maximize the benefit of the community located in or near the devastated area. Therefore, results of the baseline cohort studies are returned to each participant. In addition, statistical results of the studies are also shared with the municipalities to contribute to the improvement of the people's health in the community.

One of the major ELSI considerations regarding the participants of the cohort studies is to respect the autonomy and to protect their privacies. During the recruitment, we took care of each resident's decision to participate or not in the cohort studies independently. For this reason, we have separated the recruitment interviews not for clinical but for research setting. All informed consents (IC) were obtained by our specially trained genome medical research coordinators (GMRCs), many of whom are also certified by the Japan Society of Human Genetics (JSHG). Information of the cohort studies is de-identified and stored in our supercomputer, which is independent from the other networks. Genome sequence information is treated carefully with high-level security; all analyses using the genome sequence data are operated in the TMM supercomputer, and only analytical results without personal identification are taken out from the security area in the supercomputer.

The TMM project is the first large-scale biobank in Japan, which is set up on population cohorts and declares the sample and data sharing. As an integrated biobank, the TMM biobank is under the data- and sample-sharing policy that is not so familiar to the academia and industries. We distribute data and samples without insisting the intellectual property from us.

The TMM biobank proposes a number of novel rules as well as governing structures for better system operation and compliance with the ELSI requirements. Now, our challenge moves on to the social implementation of precision medicine, which we theorize has many ELSI concerns, especially at the interface of community medicine. To this end, we are planning to make tight relations with community medical professionals.

## Education and Training

Human resource development is necessary for the implementation of TMM project consisting of multiple activities, including community medical support, large-scale cohort study, and genome and omics analyses. As described above, our GMRCs obtained IC from all of the cohort participants via face-to-face explanation and worked for the baseline assess-

ment. Nearly 300 persons were cumulatively authorized as ToMMo GMRCs by an original training and education program, which offers more intensive training than that for GMRC certified by JSHG. Approximately half of them are still playing important roles for close communication with participants in the TMM repeat assessment center-based survey during the second phase of the TMM project.

Considering the social implementation of genomic medicine, Graduate School of Medicine at Tohoku University and Iwate Medical University, our close collaborators, have now opened genetic counseling education programs. Certificated genetic counselors and medical geneticists are working on a pilot study that returns genome analysis results of a monogenic disorder. As described above, upon the arrival of the days in which we take an advantage of PRSs or risk assessment in the clinic, we need to consider by whom and how the genetic risks of common diseases are returned for realizing precision medicine and PHC, which is the main goal of the TMM project.

In addition, it has been known that in Japan, experts of bioinformatics are mainly educated in graduate schools, whereas biobank managers and specialists of medical informatics are educated almost exclusively in on-the-job trainings. In anticipation of making more progress of genomic medicine, comprehensive education by taking advantage of TMM facilities and resources could be one of the ways to develop specialists in genomic medicine and make them available throughout Japan. In this way, it is required to develop and provide human resources not only for the TMM project but also for the social implementation of precision medicine.

## Return of Individual Genomic Results

Since the completion of the first human genome sequencing and subsequent emerging systematic catalog of human genetic variation<sup>(15)</sup>, there has been a growing expectation that the association between variations and disease susceptibility will contribute to precision medicine and PHC. We focused on three types of genetic variants that cause various disorders: rare variants in genes responsible for Mendelian diseases, variants that could modulate responses to drugs, and variants associated with the risk of complex or common diseases, including diabetes mellitus, hypertension, coronary diseases. As WGS becomes available in various research settings, obtaining individual genome information is now feasible for the application of genomics to the clinical care.

In 2013, the American College of Medical Genetics and Genomics (ACMG) published a guideline of incidental findings in exome and genome sequencing to make recommendations about responsible management of secondary findings when patients undergo exome or genome sequencing in the clinical setting<sup>(24), (25)</sup>. In the updated paper, the minimum list of the secondary findings includes 59 medically actionable genes that are recommended for return in clinical genomic sequencing.

One of the missions of our genomic cohort studies is to immediately provide the precision medicine and PHC for our cohort participants and create experience of return of individual genomic results from the research setting, which has not been systematically analyzed or operated. After extensive discussion and literature review<sup>(26)</sup>, we choose familial hypercholesterolemia (FH) as a target disease for the first pilot study of the return of genomic testing to the individual participants. FH is an actionable disease, listed in the ACMG guideline, and identifying those with FH through genomic analysis has allowed early treatment. Return of genomic results could expect outcomes such as facilitation of diagnoses, initiation of therapies at earlier ages, and more accurate risk estimation, especially for family members.

This pilot study included approximately 200 cohort participants, who finally gave informed consent after the “genetic workshop,” which contains the basic genetics, unique features of genetic information, and medical facts of FH. About 10% of the participants were positive for pathogenic or likely pathogenic SNPs in *LDLR*, *PCSK9*, and *APOB* genes. The return of the individual results was done face-to-face by clinical geneticists or certified genetic counselors. Now, we are preparing the next pilot studies regarding pharmacogenomics information and low-penetrance and adult-onset diseases, such as hereditary cancer syndrome, i.e., hereditary breast and ovarian cancer (HBOC) and/or Lynch syndrome.

## Examples of Utilizing Our Genome Data in Clinical Studies

### 1. Sequencing for rare and undiagnosed diseases

Several patients are suffering from inherited rare diseases and seeking diagnosis. To save such patients from the so-called diagnostic odyssey, clinical whole-exome sequencing (WES) or WGS was applied to identify variants causing diseases. WES or WGS was successfully applied in clinical settings owing to the decrease in sequencing cost and the improvement of sequencing technologies. As part of such efforts, the Initiative on Rare and Undiagnosed Diseases (IRUD) project is ongoing in Japan led by the Japan Agency for Medical Research and Development (AMED)<sup>(27)</sup>.

It is generally difficult to identify casual genes by sequencing only the proband, because in addition to causal variants, patients possess millions of variants irrelevant to diseases. To overcome such difficulties, it is ideal to sequence not only the probands but also their close relatives, such as parents who do not develop symptoms. Variants shared with their relatives are likely unrelated. However, it is not always possible to obtain such samples. In addition, sequencing of the relatives often results in vain, because enormous candidate variants remain even after excluding the shared variants. In such cases, ToMMo's Japanese genome reference panel plays important roles<sup>(6), (10)</sup>.

Our genome reference panel is based on whole-genome se-

quence data of the general population in Japan<sup>(10)</sup>. As a result, variants cataloged in our panel are less likely to cause severe diseases. Based on this idea, many IRUD projects have been taking advantage of our reference panel to interpret their sequencing data. Information on variant frequencies is also critical to judge whether variants are benign, and a larger number of reference panel provides more accurate frequencies for rare variants. Thus, we have been updating our reference panel from 1KJPN (1,092 samples), 2KJPN (2,049 samples), 3.5KJPN (3,554 samples), to 3.5KJPNv2 (3,552 samples due to two consent withdrawals). It should be noted that, due to technical limitations of short-read sequencing, the current reference panels do not cover large structural variants (SVs). We are expecting that long-read sequencing, which will be practical in the near future, will allow us to construct more comprehensive reference panels including SVs.

## 2. Cancer clinical sequencing

Clinical sequencing is now widely adopted as advanced medical care of cancer treatment. Information on somatic variants that contribute to tumorigenesis and tumor malignancy is valuable to determine treatment policies, such as chemotherapy. However, compared with germline variants, sequencing of cancer somatic mutations is not that easy. First, it is practically difficult to extract tumor samples without contamination of normal tissues. This is especially the case for metastatic lesion. Second, tumors often get diversification through its development, leading to genetic heterogeneity.

For these reasons, much higher sequencing coverage (depth) is necessary for cancer clinical sequencing. In general,  $\times 30$  (WGS) or  $\times 50$ -100 (WES) sequencing depth is necessary for germline variants. On the other hand,  $\times 100$  to  $\times 1,000$  sequencing depth is necessary for cancer clinical sequencing, depending on its purposes<sup>(28)</sup>. Therefore, targeted sequencing methods, such as panel sequencing focusing on cancer-related genes, are often adopted. Interestingly, it has been reported that the so-called mutation load, the number of nonsynonymous mutations, can be utilized to predict the responsiveness to cancer immunotherapy<sup>(29)</sup>. In such cases, WES or WGS is adopted to estimate the accurate mutation load by analyzing comprehensive datasets of somatic mutations.

Examination of germline mutations is also important in cancer genomics, because 5% to 10% of cancers are believed to be related to inherited variants in tumor suppressor genes, such as cell cycle-related genes and DNA repair genes<sup>(30)</sup>. To estimate the frequency of such cancer risk variants, ToMMo's reference panel will be helpful. For instance, we previously summarized the frequencies of pathological variants in the 2KJPN, for which the ACMG recommends returning of the results<sup>(31)</sup>. According to this report, the 2KJPN reference panel includes five reported pathogenic variants in *BRCA1* and *BRCA2* genes whose germline mutations predispose to breast and ovarian cancer. The data also allows us to estimate the population frequencies of these susceptible variants. In this

way, the reference panel of the general population is a powerful tool to obtain the fundamental knowledge necessary for genomic medicine.

## Conclusions

To realize precision medicine and PHC, integration of the genomic and omics data with lifestyle data is indispensable for the improvement of disease risk assessment and understanding of the biological processes of diseases for prevention, early diagnosis, and proper managements. We believe that prospective cohort studies in the TMM project will provide elaborate genomic and omics data platforms that actualize the genomic medicine. In the 21st century, many scientists pay great attention to precision medicine and PHC. The goals of our integrated biobank are to promote healthy lifestyles and to prevent and develop treatment methods suitable for each person's constitution.

## Article Information

### Conflicts of Interest

None

### Sources of Funding

This work was supported in part by Tohoku Medical Megabank Project from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and AMED (grant numbers JP18km0105001, JP18km0105002), Platform Program for Promotion of Genome Medicine (P3GM; grant number JP18km0405001), Advanced Genome Research and Bioinformatics Study to Facilitate Medical Innovation in P3GM (GRIFIN; grant number JP18km0405203), and Center of Innovation Program from Japan Science and Technology Agency (JST).

### Acknowledgement

We would like to thank all the volunteers who participated in this study. We would also like to acknowledge all those associated with the study, especially Ms. Miho Kuriki for creation of artwork. The member list is available at the following website: <http://www.megabank.tohoku.ac.jp/english/a181201/>. We would also like to thank the Iwate Medical University Iwate Tohoku Medical Megabank Organization (IMM) for collaboration.

### Author Contributions

Nobuo Fuse and Mika Sakurai-Yageta contributed equally to this work.

### Approval by Ethical Committee

The project protocols were reviewed and approved by the Ethics Committee of Tohoku University Graduate School of Medicine and by the Ethics Committee of Iwate Medical Uni-



versity.

## References

- Ishigaki A, Higashi H, Sakamoto T, et al. The Great East-Japan Earthquake and devastating tsunami: an update and lessons from the past Great Earthquakes in Japan since 1923. *Tohoku J Exp Med*. 2013;229(4):287-99.
- Nishi N, Yoshimura E, Ishikawa-Takata K, et al. Relationship of living conditions with dietary patterns among survivors of the great East Japan earthquake. *J Epidemiol*. 2013;23(5):376-81.
- Murakami H, Yoshimura E, Ishikawa-Takata K, et al. The longitudinal change in physical activity among Great East Japan Earthquake victims living in temporary housing. *Nihon Koshu Eisei Zasshi*. 2014;61(2):86-92.
- Kuriyama S, Yaegashi N, Nagami F, et al. The Tohoku Medical Megabank Project: Design and Mission. *J Epidemiol*. 2016;26(9):493-511.
- Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-53.
- Nagasaki M, Yasuda J, Katsuoka F, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*. 2015;6:8018.
- Kawai Y, Mimori T, Kojima K, et al. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *J Hum Genet*. 2015;60(10):581-7.
- Tadaka S, Saigusa D, Motoike IN, et al. jMorp: Japanese Multi Omics Reference Panel. *Nucleic Acids Res*. 2018;46(D1):D551-7.
- Pearson H. Massive UK baby study cancelled. *Nature*. 2015;526(7575):620-1.
- Yasuda J, Kinoshita K, Katsuoka F, et al. Genome analyses for the Tohoku Medical Megabank Project toward establishment of personalized healthcare. *J Biochem*. 2019;165(2):139-58.
- Koshiha S, Motoike I, Kojima K, et al. The structural origin of metabolic quantitative diversity. *Sci Rep*. 2016;6:31463.
- Koshiha S, Motoike I, Saigusa D, et al. Omics research project on prospective cohort studies from the Tohoku Medical Megabank Project. *Genes Cells*. 2018;23(6):406-17.
- Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-9.
- International HapMap C. The International HapMap Project. *Nature*. 2003;426(6968):789-96.
- The 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
- Wray NR, Yang J, Hayes BJ, et al. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*. 2013;14(7):507-15.
- Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet*. 2006;7(10):812-20.
- Ueki M, Tamiya G; Alzheimer's Disease Neuroimaging Initiative. Smooth-threshold multivariate genetic prediction with unbiased model selection. *Genet Epidemiol*. 2016;40(3):233-43.
- Hachiya T, Kamatani Y, Takahashi A, et al. Genetic predisposition to ischemic stroke: A polygenic risk score. *Stroke*. 2017;48(2):253-8.
- Editorial. GWAS to the people. *Nat Med*. 2018;24(10):1483.
- Xu P, Gunsolley J. Application of metagenomics in understanding oral health and disease. *Virulence*. 2014;5(3):424-32.
- Leishman SJ, Ford PJ, Do HL, et al. Periodontal pathogen load and increased antibody response to heat shock protein 60 in patients with cardiovascular disease. *J Clin Periodontol*. 2012;39(10):923-30.
- Zheng W, Zhang Z, Liu C, et al. Metagenomic sequencing reveals altered metabolic pathways in the oral microbiota of sailors during a long sea voyage. *Sci Rep*. 2015;5:9131.
- Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*. 2013;15(7):565-74.
- Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19(2):249-55.
- Aizawa Y, Kobayashi T, Kawame H. [Review of present state and challenges regarding return of genomic results from genome-wide analysis in Japan: Consideration of American College of Medical Genetics and Genomics (ACMG) recommendation for reporting incidental findings]. *Jpn J Genet Counsel*. 2016;37(3):105-26. Japanese.
- Adachi T, Kawamura K, Furusawa Y, et al. Japan's initiative on rare and undiagnosed diseases (IRUD): towards an end to the diagnostic odyssey. *Eur J Hum Genet*. 2017;25(9):1025-8.
- Ajay SS, Parker SC, Abaan HO, et al. Accurate and comprehensive sequencing of personal genomes. *Genome Res*. 2011;21(9):1498-505.
- Lyu GY, Yeh YH, Yeh YC, et al. Mutation load estimation model as a predictor of the response to cancer immunotherapy. *NPJ Genom Med*. 2018;3:12.
- Romero-Laorden N, Castro E. Inherited mutations in DNA repair genes and cancer risk. *Curr Probl Cancer*. 2017;41(4):251-64.
- Yamaguchi-Kabata Y, Yasuda J, Tanabe O, et al. Evaluation of reported pathogenic variants and their frequencies in a Japanese population based on a whole-genome reference panel of 2049 individuals. *J Hum Genet*. 2018;63(2):213-30.

JMA Journal is an Open Access journal distributed under the Creative Commons Attribution 4.0 International License. To view the details of this license, please visit (<http://creativecommons.org/licenses/by/4.0/>).