Twin Research and Human Genetics
Volume 16 ▌ Number 1 ▌ pp. 241–245 ▌ © The Authors 2012 ▌ doi:10.1017/thg.2012.130

The Healthy Twin Study, Korea Updates: Resources for Omics and Genome Epidemiology Studies

Bayasgalan Gombojav,[1] Yun-Mi Song,[2] Kayoung Lee,[3] Sarah Yang,[4] Minjung Kho,[4] Yong-Chul Hwang,[4] Gwangpyo Ko,[5,6] and Joohon Sung[1,4]

[1]Institute of Environment and Health, School of Public Health, Seoul National University, Seoul, Korea
[2]Department of Family Medicine, Samsung Medical Center, and Center for Clinical Research, Samsung Biomedical Research Institute, Sung Kyun Kwan University School of Medicine, Seoul, Korea
[3]Department of Family Medicine, Busan Paik Hospital, Inje University College of Medicine, Gimhae, Korea
[4]Complex Disease and Genome Epidemiology Branch, Department of Epidemiology, School of Public Health, Seoul National University, Seoul, Korea
[5]Environmental Microbiology Branch, Department of Environmental Health, School of Public Health, Seoul National University, Seoul, Korea
[6]Microbial Systems & Communities, Genome Sequence and Analysis Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

The Healthy Twin Study, Korea (HT) is an ongoing multi-center cohort study that was initiated in 2005, based on a nation-wide twin and family database. Since its inception, the HT has recruited 815 pairs of adult twins and a total of 3,690 individual twins and their families as of July 2012. Here we summarize updates since the previous report in 2006. Besides the increase in size, the HT has been enriched in several aspects: a biobank was constructed for ongoing and future omics studies; and genome-wide single nucleotide polymorphism markers (Affymetrix GeneChip version 6.0, 1 M probes) have been analyzed for 2,200 individuals, which enabled gene identification studies for measured phenotypes. In addition, longitudinal study protocols were established through the HT and a second wave survey was finished in 2010 with >70% follow-up rate. The parallel genome research projects were recently launched, which would expedite multi-omics studies maximizing the twin potentials such as metagenomics and epigenetics studies, and endow us with resources for recruiting more participants. We submit this report to share updates and research opportunities from the HT.

■ **Keywords:** twin research, twin registry, cohort study, genomics

The Healthy Twin Study, Korea (HT) is an ongoing multi-center cohort study, which was initiated in 2005 from a nation-wide twin and family database (Sung et al., 2002, 2006b). Since its inception, the HT has recruited 815 pairs of adult twins (a total of 3,690 individuals of twins and their families). The protocols and measurements are described in detail in the previous report (Sung et al., 2006a). In brief, the HT is recruiting adult like-sex twins through a national database, together with their first-degree family members, with no ascertainment by specific disease status. Extended questionnaires and health examinations have been provided upon recruitment and follow-up.

Since 2006, when the first report on study design of the HT was published, the HT has been enriched in several aspects: the size of the registry has increased and now includes 3,690 participants as of July 2012; construction of a biobank has been successful and is tailored for multiple 'omics' studies; protocols for follow-up have been established and the second wave survey was conducted in 2011, with a greater than 70% follow-up rate; several additional genome research projects have also been started and will allow multi-dimensional omics data to be generated and the ability to recruit more participants with multiple funding sources. We expect that this update provides information of potential interest both inside and outside of the twin research community.

RECEIVED 31 August 2012; ACCEPTED 29 October 2012. First published online 7 December 2012.

ADDRESS FOR CORRESPONDENCE: Joohon Sung, Complex Disease and Genome Epidemiology Branch, Department of Epidemiology and Institute of Environment and Health, School of Public Health Seoul National University, 1st Kwanak-Ro, Kwanak-Gu, Seoul, 151–742, Korea. E-mail: jsung@snu.ac.kr

TABLE 1

Age, Sex, and Zygosity Distribution of the Healthy Twin Study Participants as of July, 2012, Number of Pairs (for Twins) and Number of Individuals (for Singletons and Total Numbers)

	Twins (number of pairs)										Singleton family members (individuals)			
	MZ* pairs			DZ* pairs			XZ† pairs							
Age group	Men	Women	MZ sum	Men	Women	DZ sum	Men	Women	XZ sum	Total twin pairs	Men	Women	Subtotal of family	Total number of individuals
<29	34	41	75	11	1	12	0	1	1	88	137	148	285	461
30–39	123	211	334	38	43	81	5	9	14	429	182	249	431	1,289
40–49	62	111	173	12	22	34	2	0	2	209	144	233	377	795
50–59	31	32	63	11	6	17	2	2	4	84	151	257	408	576
60–69	2	2	4	0	0	0	0	0	0	4	196	234	430	438
70–79	0	1	1	0	0	0	0	0	0	1	53	69	122	124
80+	0	0	0	0	0	0	0	0	0	0	4	3	7	7
Subtotal	252	398	650	72	72	144	9	12	21	815	867	1,193	2,060	3,690

Note: MZ = monozygotic twin; DZ = dizygotic twin; XZ = zygosity undetermined twin.*Zygosity estimation was based on genotype between 2005 and 2009, and based on questionnaire only since 2010. †Zygosity of 'XZ' was assigned either in case questionnaire-based zygosity survey showed discrepancy between the co-twins (i.e., MZ vs. DZ) or both co-twins fell in the category of ambiguous zygosity.

Updates on Participants

Through continued recruitment from a nation-wide database the Korean Twin Registry using the same procedures described in previous report (Sung et al., 2002), we have recruited a total of 3,690 individuals as of July 2012: 45% are like-sex twins, and the rest are their family members. An outline of current participants is described in Table 1. Because twin pairs are also allowed to join without their families, and larger families are encouraged to participate in the study, the size distribution of families is bimodal, with a peak at a family size of four. We believe the protocol of recruiting families along with twins has made the participation of monozygotic (MZ) easier. The HT has more MZ twins (78%) than dizygotic twins (DZ) and more females (61%). Excluding opposite-sex twins by design resulted in the excess of MZs. Zygosity determination was performed using a questionnaire developed by the authors (Song et al., 2010) or through genetic markers, which are ambiguous by the questionnaire survey (Christiansen et al., 2003; Song et al., 2010).

Establishment of Biobank for Multi-Omics Studies

The HT was designed initially to enable omics studies, and so establishing a biobank was a priority. For each participant, genomic DNA was extracted and aliquoted; buffy coat fraction, enriched for white blood cells and platelets, was either treated for RNAse inhibitor (RNA later, until 2006), or snap frozen in liquid nitrogen (–180°C; since 2007). Epstein-Barr virus-transfected lymphoblastic B-cell lines have been generated for about 65% of participants as a semi-inexhaustible resource of DNA; plasma and serum are centrifuged within 90 minutes of collection, and two vials of serum and plasma are immediately transferred to a portable liquid nitrogen tank to meet proteomics quality standards. Additionally, 12-hour urine samples are collected along with the information of collection time and total volume, which will be useful for analyzing metabolites and biomarkers of exposure. One vial of urine sample for each participant is also kept in a deep freezer (–80°C), other samples are stored at –25°C.

In 2010, a microbial study was initiated and additional participants were recruited. In this microbial study involving twins and their parents and siblings, samples collected included stool samples (250 mL stool box), sputum, cervical smears (from women who took a Papanicolaou test), two smear samples from skin (dorsal surface or arm and back), and for the oral cavity, supra- and sub-gingival swab samples with mouthwash fluid after washing with clean water. For microbial samples, an aseptic cotton ball or toothpick (oral cavity) were collected together with microbial samples to serve as controls. Separate informed consent was obtained regarding storage of specimens, duration of storage, and use of information and scientific or commercial products that would be generated from those biospecimens. Table 2 describes the summary information of the biobank.

Genotyping and Quality Control (QC) of Genetic Markers

Genomic DNA was extracted from venous blood samples drawn on all participants at their health examinations, and genotyping was performed in 2009 with the Affymetrix Genome-Wide Human SNP Array version 6.0. A conventional QC procedure for dense short nucleotide polymorphism (SNP) markers was carried out (WTCCC, 2007) and additional extensive marker cleaning was performed using familial relationships. In addition, the following SNPs were excluded: duplicated (3,011 SNPs); Hardy–Weinberg disequilibrium $p < .001$ or minor allele frequency <0.1 (288,426 SNPs); genotype missing rate >0.05 (4,227 SNPs); Mendelian inconsistency in >3 families (11,456 SNPs); and

https://doi.org/10.1017/thg.2012.130 Published online by Cambridge University Press

TABLE 2

Summary Information of the Biobank

Type of biospecimens	Volume or unit per specimen	Number of specimens per individual	Storage temperature	Possible Omics application	Notes
Blood components					Every participant
Centrifuge and aliquoting was done within 2 hours from sampling, and buffy coats, sera, and plasma were quick-freezed in LN2					
DNA extract	10 μ per vial (x2) and rest (1)	>3	Deep freezer (−80°C)	G, Eg	
Lymphoblastic B-Cell line (EBV transfected)	>7 × 10^6 cells per 1.0 mL	>2	LN2 tank (−180°C)	G, Tr	Stopped since 2010
Buffy coats	0.5 mL	>1	LN2 tank (−180°C)	G, Eg, Tr	Anti-RNAse treated until 2007
Serum	1.0 mL	>3	LN2 tank (−180°C)	Pr, Metab	
Plasma	1.0 mL	>3	LN2 tank (−180°C)	Pr, Metab	
Packed red blood cell (1)	1.0 mL	1	Deep freezer (−80°C)	Exposure assessment	
Other biospecimens					Every participant
Most urine samples were collected as half-day protocol, but those specimens with incomplete volume-time are considered as spot urine					
Half-day urine	12 mL tube	>3	Freezer (−25°C)	Metab	
Half-day urine	1.0 mL	1>	Deep freezer (−80°C)		
Microbial sample					>350 persons
Stool	250 mL bo X15 mL box	1 >1	Deep freezer (−80°C)	Mic	
Cervical smear	DNA extraction	1	Deep freezer (−80°C)	Mic	
Skin swap	Cotton ball	1	Deep freezer (−80°C)	Mic	
Sputum	12 mL box	1	Deep freezer (−80°C)	Mic	
Oral cavity swap	Cotton ball	1	Deep freezer (−80°C)	Mic	

Note: G = genomics; Eg = epigenomics; Tr = transcriptomics; Pr = proteomics; Metab = metabolomics; Mic = microbiome (= metagenomics).

non-Mendelian multi-marker inconsistency in >3 families (47,594 SNPs). These exclusions reduced the total number of SNPs from 891,873 to 537,159.

Combined Asian HapMap Reference and Imputation of SNP Markers

To facilitate collaborative studies involving genome-wide association analyses, a marker imputation was carried out to increase compatibilities with other SNP marker sets. We built a new reference marker set using Asian HapMap3 data (release 2) with 1.39 million SNP markers and Korean HapMap data consisting of 90 unrelated Koreans with 1.66 million SNP markers (http://www.khapmap.org). HapMap3 Asian panel was used to phase the Korean HapMap, resulting in 1.39 million markers of 260 persons. Family-based SNP marker imputation was performed in three steps; first, each individual was treated as unrelated and a conventional method was applied using IMPUTE2 (Marchini & Howie, 2010); then Mendelian incompatible markers within the family were deleted for all family members; next, only those missing markers of the founders in each family were imputed again using IMPUTE2; finally, family-wise imputation was performed using BEAGLE, where the genetic markers of the founders were used as a reference in each family to impute the non-founders' missing information. SNPs which had R^2 (ratio of the variance of imputed genotypes to the binomial variance) <0.03 were excluded, resulting in a total of 1,387,466 SNPs. The mean (*SD*) of the imputation score (r^2) was 0.997 (0.028).

Protocols for Longitudinal Study

The HT started follow-up survey in 2008. Participants are invited to take full health examination and questionnaire survey every 3 years. A third wave follow-up began in 2012. In the second wave, 1,848 participants were re-examined out of 2,602 target individuals (72%; Table 3). Most health examinations and questionnaire-based surveys are repeated in every wave, with the exception of some tests which are measured only after 5-year intervals, such as lung function or carotid artery Doppler scan. Some measurements such as whole body dual-energy X-ray absorptiometry and echocardiogram are only taken at the initial examination. Medical history between the health examinations are too (www.twinkorea.org for detailed follow-up protocols). New participants consisted of the non-participant members of existing families or those recruited from new research projects such as the Korean Microbiome study.

Multiple Omics Study Projects

The HT was initiated by the support of The Genome Research Center of Center for Disease Control, Korea. The Korean Microbiome Project and other research projects are being conducted as parallel projects, which enrich the data, information, and participants of the HT.

https://doi.org/10.1017/thg.2012.130 Published online by Cambridge University Press

TABLE 3

Summary of the Follow-Up Study and New Cases of Hypertension and Overweight Detected During the Second Wave Survey Between 2008 and 2011

	Twin									Families									Total		
	Men			Women			Subtotal			Men			Women			Subtotal					
Age group	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight	Total follow-up	Hypertension	Overweight
<29	4		2	31	1		35	1	2	47		4	49		3	96		7	131	1	9
30–39	185		11	313	2	12	498	2	23	82	3	7	126		8	208	3	15	706	5	38
40–49	71		4	161	4	9	232	4	13	75	2	6	117	3	5	192	5	11	424	9	24
50–59	42	1	3	45	3		87	4	3	61		6	130		8	191		14	278	4	17
60–69	2	5	1	2	1		4	6	1	106		2	134		7	240		9	244	6	10
70–79				2	1		2	1		21		2	39		3	60		5	62	1	5
80+										3						3			3		
Subtotal	304	6	21	554	12	21	858	18	42	395	5	27	595	3	34	990	8	61	1,848	26	103

The Korean Microbiome Project

The Korean Microbiome project is analogous to the Human Microbiome Project (Gevers et al., 2012), which aims to characterize the microbial communities that live on the human body, and to examine associations between microbial diversity and susceptibilities of human disease (Gevers et al., 2012). The Korean Microbiome Project was designed to recruit mainly twins. It is well known that a twin study design particularly suits a microbiome study, because discordance in microbial profile among MZ pairs will provide reliable evidence of associations by canceling out noise from genomic DNA sequences or unmeasured environmental factors. For more than 200 stool samples of twins, the V2 and V3 regions of bacteria-specific 16S rRNA genes were amplified and sequenced by using the 454 Life Sciences FLX Titanium (Roche, Indianapolis, IN, USA) or Illumina HiSeq (Illumina, San Diego, CA, USA). After removing low-quality sequences (quality score < 25) phylogenetic analyses and taxa allocation were done using Quantitative Insights Into Microbial Ecology (http://qiime.sourceforge.net; Caporaso et al., 2010). Association studies between microbial profile and obesity, metabolic syndrome are being conducted.

Other Parallel Projects

Recently, the National Research Foundation, Korea commenced a new project involving twins. The Global Research Network program (GRN, 2011–2014), a pilot phase program, accepted a grant to support international collaboration between twin registers. The GRN aims to: (1) facilitate a global twin registry Network; (2) collect obesity-discordant and cardiovascular disease-discordant twin pairs, through the network; (3) identify risk factors and complications of both underweight and obesity using the 'normal' weight co-twin as a control (co-twin–control study). Despite the unique strength of studies involving discordant twins, the single most important barrier of co-twin–control studies is the scarcity of them. Identical twins comprise 0.3–0.6% of populations and only a small fraction of them are discordant for the phenotypes of interest. International collaboration will enable the researchers to recruit discordant twin pairs. Some new twin pairs, particularly disease- or health status-discordant MZ twins will be recruited from this program.

More support came from the Next Generation Personalized Medicine Project, Korea (PGM21). This program aims to discover genetic variants associated with common important diseases and conditions, and to apply the genetic information to generate preventive strategies. We are conducting a wide range of genome-wide association studies with not only disease outcomes, but with disease prediction models which can be applied to preventive measures. The PGM21 (2012–2015) will support the analysis of existing data, and will support further recruitment of participants and genotyping, which will be necessary for the replication of findings.

Conclusion

The HT has expanded considerably since the previous report of 2006. With the growing size, modern resources for omics studies, and increase in multi-dimensional information, the HT will be able to serve as valuable resources for twin research, suitable for common disease and risk factor genetic epidemiology studies.

Acknowledgments

This study was supported by the National Genome Research Institute, Korea, National Institute of Health research contract (budgets 2011E7101100, 2012E7100200), National

https://doi.org/10.1017/thg.2012.130 Published online by Cambridge University Press

Research Foundation of Korea (NRF 2011-220-E00006; NRF 2010-0029113; NRF 2012K2A1A2032536; and NRF 2010-0025814), PGM21 (A111218-12-GM02). SY, MJK, and YCH were supported by BK21 program. The views expressed in this article are those of the authors and not necessarily any funding body.

References

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *5*, 335–336.

Christiansen, L., Frederiksen, H., Schousboe, K., Skytthe, A., von Wurmb-Schwark, N., Christensen, K., & Kyvik, K. (2003). Age-and sex-differences in the validity of questionnaire-based zygosity in twins. *Twin Research*, *4*, 275–278.

Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., . . . Huttenhower, C. (2012). The human microbiome project: A community resource for the healthy human microbiome. *Plos Biology*, *8*, e1001377.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *7*, 499–511.

Song, Y. M., Lee, D., Lee, M. K., Lee, K., Lee, H. J., Hong, E. J., . . . Sung, J. (2010). Validity of the zygosity questionnaire and characteristics of zygosity — Misdiagnosed twin pairs in the Healthy Twin Study of Korea. *Twin Research and Human Genetics*, *3*, 223–230.

Sung, J., Cho, S. H., Cho, S. I., Duffy, D. L., Kim, J. H., Kim, H., . . . Park, S. K. (2002). The Korean Twin Registry-methods, current stage, and interim results. *Twin Research*, *5*, 394–400.

Sung, J., Cho, S. I., Lee, K., Ha, M., Choi, E. Y., Choi, J. S., . . . Song, Y. M. (2006a). Healthy Twin: A twin-family study of Korea — Protocols and current status. *Twin Research and Human Genetics*, *6*, 844–848.

Sung, J., Cho, S. I., Song, Y. M., Lee, K., Choi, E. Y., Ha, M., . . . Kimm, K. (2006b). Do we need more twin studies? The Healthy Twin Study, Korea. *International Journal of Epidemiology*, *2*, 488–490.

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *7145*, 661–678.

https://doi.org/10.1017/thg.2012.130 Published online by Cambridge University Press