DATA REPORT

iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation

Shohei Komaki1,8, Yuh Shiwa1,2,3,8, Ryohei Furukawa1, Tsuyoshi Hachiya1, Hideki Ohmomo1,2, Ryo Otomo1,2, Mamoru Satoh1,2, Jiro Hitomi4,5, Kenji Sobue6, Makoto Sasaki5,7 and Atsushi Shimizu1

We launched an integrative multi-omics database, iMETHYL (http://imethyl.iwate-megabank.org). iMETHYL provides whole-DNA methylation (~24 million autosomal CpG sites), whole-genome (~9 million single-nucleotide variants), and whole-transcriptome (~14 000 genes) data for CD4+ T-lymphocytes, monocytes, and neutrophils collected from approximately 100 subjects. These data were obtained from whole-genome bisulfite sequencing, whole-genome sequencing, and whole-transcriptome sequencing, making iMETHYL a comprehensive database.

Human Genome Variation (2018) 5, 18008; doi:10.1038/hgv.2018.8; published online 29 March 2018

DNA methylation (DNAm) has a critical role in regulating gene expression. Recent epigenome-wide association studies in humans have revealed that locus-specific DNAm signatures are associated with susceptibility to different environmental exposures, intermediate phenotypes, and diseases.1,2 Hence, locus-specific DNAm signatures are potential biomarkers in the era of precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicine.3 We recently found that CpG sites with large interindividual DNAm variation are more likely to be potential biomarkers,4 suggesting that a database of interindividual DNAm variation would be useful to determine target regions for future precision medicin

1Division of Biomedical Information Analysis, Iwate Medical University, Shiwa, Iwate, Japan; 2Division of Biobank and Data Management, Iwate Medical University, Shiwa, Iwate, Japan; 3Laboratory of Bioinformatics, Department of Molecular Microbiology, Faculty of Life Sciences, Tokyo University of Agriculture, Setagaya, Tokyo, Japan; 4Department of Anatomy, School of Medicine, Institute for Biomedical Sciences, Iwate Medical University, Shiwa, Iwate, Japan; 5Iwate Tohoku Medical Megabank Organization, Iwate Medical University, Shiwa, Iwate, Japan; 6Iwate Medical University, Morioka, Iwate, Japan; 7Division of Ultrahigh Field MRI, Institute for Biomedical Sciences, Iwate Medical University, Shiwa, Iwate, Japan.

Correspondence: Atsushi Shimizu (ashimizu@iwate-med.ac.jp)

Received 31 October 2017; revised 31 December 2017; accepted 15 January 2018
frequencies for ∼9 million autosomal SNVs (Table 1). Statistics regarding age, sex, and database profiles used in iMETHYL are presented in Table 1. Furthermore, genomic annotation tracks, such as gene models, repetitive elements, CpG islands, and microarray probes, are available in the iMETHYL browser (Table 2).

Table 1. Demographic and profile statistics of iMETHYL

| Demographic characteristics of subjects | Monocytes | CD4+ T cells | Neutrophils |
|-----------------------------------------|-----------|--------------|-------------|
| N                                      | 102a      | 102a         | 94          |
| Males, N (%)                           | 48 (47.1) | 49 (48.0)    | 48 (51.1)   |
| Median age (range), years              | 62.5 (35–75) | 62.0 (35–75) | 58.0 (24–81) |
| DNAm profiles                          |           |              |             |
| No. of autosomal CpGs^c                | 23,941,821| 24,037,518   | 25,483,031  |
| Gene expression profiles               |           |              |             |
| No. of sequencing reads^b              | 33,917,157 ± 3,153,528 | 35,175,996 ± 1,275,575 | 47,040,140 ± 6,289,540 |
| No. of genes^d                         | 16,282    | 18,299       | 14,534      |
| SNV profiles                           |           |              |             |
| No. of SNVs^e                          | 8,945,669 | 8,951,822    | 8,792,880   |

Abbreviations: DNAm, DNA methylation; SNV, single-nucleotide variant. aBoth cell types were obtained from the same 95 individuals out of a cohort of 102. bAverage ± standard deviation. cCpGs that were retained in ≥50% of subjects for each cell type. dGenes that were expressed with a fragments per kilobase of exon per million mapped fragments ≥0.1 in ≥50% of subjects for each cell type. eSNVs with a minor allele count > 1.

Table 2. List of available tracks in iMETHYL

| Track name        | Description                                           | Source |
|-------------------|-------------------------------------------------------|--------|
| IMM_CpG_CD4T      | Information for each CpG site of CD4T                 | Ref. 4 |
| IMM_CpG_CD4T_avg  | Average DNAm level of each CpG site of CD4T           | Ref. 4 |
| IMM_CpG_CD4T_sd   | DNAm variations of each CpG site of CD4T measured by SD| Ref. 4 |
| IMM_CpG_CD4T_RI   | DNAm variations of each CpG site of CD4T measured by RI| Ref. 4 |
| IMM_CpG_Mono      | Information for each CpG site of monocytes            | Ref. 4 |
| IMM_CpG_Mono_avg  | Average DNAm level of each CpG site of monocytes      | Ref. 4 |
| IMM_CpG_Mono_sd   | DNAm variations of each CpG site of monocytes measured by SD| Ref. 4 |
| IMM_CpG_Mono_RI   | DNAm variations of each CpG site of monocytes measured by RI| Ref. 4 |
| IMM_CpG_Neu       | Information for each CpG site of neutrophils          | This study |
| IMM_CpG_Neu_avg   | Average DNAm level of each CpG site of neutrophils    | This study |
| IMM_CpG_Neu_sd    | DNAm variations of each CpG site of neutrophils measured by SD| This study |
| IMM_CpG_Neu_RI    | DNAm variations of each CpG site of neutrophils measured by RI| This study |
| IMM_FPKM_CD4T     | FPKM values of each transcript of CD4T                | Ref. 4 |
| IMM_FPKM_Mono     | FPKM values of each transcript of monocytes           | Ref. 4 |
| IMM_FPKM_Neu      | FPKM values of each transcript of neutrophils         | Ref. 4 |
| IMM_SNV_CD4T      | Information for each SNV of CD4T                      | Ref. 4 |
| IMM_SNV_Mono      | Information for each SNV of monocytes                 | Ref. 4 |
| IMM_SNV_Neu       | Information for each SNV of neutrophils               | Ref. 4 |
| Reference sequence | Human genome hg19/GRCh37 sequence                     | UCSC genome browser |
| RepeatMasker      | Repetitive elements                                    | UCSC genome browser |
| CpGIslandsExt     | CpG island locations                                   | UCSC genome browser |
| HM450             | Probe information for Illumina HumanMethylation450    | UCSC genome browser |
| gencode_v19       | Information of genes obtained from GENCODE version 19| GENCODE |
| gencode_v19_trs   | Information of transcripts obtained from GENCODE version 19| GENCODE |

Abbreviations: CD4T, CD4+ T-lymphocyte; DNAm, DNA methylation; FPKM, fragments per kilobase of exon per million fragments mapped; RI, reference interval; SD, standard deviation; SNV, single-nucleotide variant.

iMETHYL was developed to provide an informative, easy-to-use resource that enables investigators to explore DNAm levels and the variability of potential biomarkers identified by epigenome-wide association studies or candidate gene approach studies. From the iMETHYL browser, regions of interest can be specified using gene symbols (GENCODE release 19), dbSNP ID, DNA methylation array probe ID, and genomic positions. The genome browser provides graphical views of genomic annotations and the average methylation level and variability (SD and RI) of each CpG site in each of the three human cell types (Figure 1a). In addition, tracks for the average expression level and SD of each gene for each cell type and allele frequencies of each SNV within 102 (CD4T), 102 (monocytes), and 94 (neutrophils) subjects are provided.

In the example shown in Figure 1a, the iMETHYL genome browser showed different tracks in the region flanking cg05575921, which is a DNAm biomarker for tobacco smoking12,13 located in the aryl-hydrocarbon receptor repressor (AHRR) gene. This DNAm biomarker is markedly demethylated in
current smokers. Using iMETHYL, the average methylation level and variability of each CpG site in the three cell types (CD4T, monocytes, and neutrophils) are shown, and by selecting the bar in the CpG tracks, histograms of DNAm levels at this CpG site for each cell type appear in pop-up windows (Figure 1b–d). iMETHYL is also useful for investigating cell-type-specific DNAm variability. In the CpG site shown in Figure 1, the DNAm levels in CD4T were hypermethylated with a narrow distribution (Figure 1b), whereas broader distributions of DNAm levels were found in monocytes and neutrophils (Figure 1c and d).

Furthermore, investigators can use the browser to explore variability in gene expression and SNVs. For example, upon selecting the bar shown in the fragments per kilobase of exons per million mapped fragment tracks, a histogram of gene expression levels appears in the pop-up window. In addition, the average expression level and SD for each gene are shown. This information provides important clues into the functional relevance of known or putative DNAm biomarkers.

Data on the mean and variation of the DNAm level of each CpG site can be downloaded from the iMETHYL website so that users can find CpG sites of their own interest based on the DNAm level and variation or differences between cell types.

In summary, we constructed a public database, iMETHYL, that provides a reference for human DNAm variation. iMETHYL is the first database featuring interindividual DNAm variation based on...
high-coverage whole-genome bisulfite sequencing using purified CD4T, monocytes, and neutrophils. Because the data were obtained from apparently healthy subjects, the multi-omics genomic data provided by iMETHYL can be used as a reference control. Investigators can examine DNAm variation, gene expression, and SNVs at any specific region of the human genome, which can enable the identification of variable regions in the population to design assay probes for microarrays or targeted sequencing. iMETHYL provides multi-omics data for three different cell types to the scientific community. The iMETHYL browser will be a useful resource not only for researchers specializing in epigenomics but also for those interested in the interactive analysis of DNA methylation, gene expression, and genomic variation.

ACKNOWLEDGEMENTS
This work was supported by the Tohoku Medical Megabank Project (Special Account for Reconstruction from the Great East Japan Earthquake) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Agency for Medical Research and Development (AMED). We thank the members of the Iwate Tohoku Medical Megabank Organization of Iwate Medical University (IMM) and the Tohoku Medical Megabank Organization of Tohoku University (ToMMo) for their encouragement and support. We especially acknowledge Dr. Fumiki Katsuoka, Professor Jun Yasuda, and Professor Masao Nagasaki for their contributions to whole-genome sequencing and analysis. We are grateful to the Tohoku Medical Megabank Project participants.

COMPETING INTERESTS
The authors declare no conflict of interest.

PUBLISHER’S NOTE
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES
1 Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet 2011; 12: 529–541.
2 Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. Nat Rev Genet 2013; 14: 585–594.
3 Andersen AM, Dogan MV, Beach SR, Philibert RA. Current and future prospects for epigenetic biomarkers of substance use disorders. Genes (Basel) 2015; 6: 991–1022.
4 Hachiya T. et al. Genome-wide identification of inter-individually variable DNA methylation sites improves the efficacy of epigenetic association studies. NPJ Genome Med 2017; 2: 11.
5 Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nat Rev Genet 2016; 17: 319–332.
6 Reinius LE. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE 2012; 7: e41361.
7 Chatteryee A. et al. Genome-wide DNA methylation map of human neutrophils reveals widespread inter-individual epigenetic variation. Sci Rep 2015; 5: 17328.
8 Shen H, Qiu C, Li J, Tian Q, Deng HW. Characterization of the DNA methylome and its interindividual variation in human peripheral blood monocytes. Epigenomics 2013; 5: 255–269.
9 Funakawa R. et al. Intraindividual dynamics of transcriptome and genome-wide stability of DNA methylation. Sci Rep 2016; 6: 26424.
10 Harrow J. et al. ENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012; 22: 1760–1774.
11 Buels. R et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016; 17: 66.
12 Tsaprouni LG. et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics 2014; 9: 1382–1396.
13 Zeilinger S. et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS ONE 2013; 8: e63812.

Supplemental Information for this article can be found on the Human Genome Variation website (http://www.nature.com/hgv).