iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing

Yumi Yamaguchi-Kabata1,2, Naoki Nariai1,5, Yosuke Kawaji1,2, Yukuto Sato1,2, Kaname Kojima1,2,3, Minoru Tateno1, Fumiki Katsuoka1,2, Jun Yasuda1,2, Masayuki Yamamoto1,2 and Masao Nagasaki1,2,3,4

The integrative Japanese Genome Variation Database (iJGVD; http://ijgvd.megabank.tohoku.ac.jp/) provides genomic variation data detected by whole-genome sequencing (WGS) of Japanese individuals. Specifically, the database contains variants detected by WGS of 1,070 individuals who participated in a genome cohort study of the Tohoku Medical Megabank Project. In the first release, iJGVD includes > 4,300,000 autosomal single nucleotide variants (SNVs) whose minor allele frequencies are > 5.0%.

Human Genome Variation (2015) 2, 15050; doi:10.1038/hgv.2015.50; published online 26 November 2015
Figure 1. Schema of the systems and graphical user interfaces of iJGVD. (a) Schematic diagram of the iJGVD systems. (b–d) Graphical user interfaces for iJGVD. (b) SNV searches are initiated at the top page by specifying a gene, dbSNP ID, or genomic region. (c) SNV allele frequencies are displayed in a table, and rs671 is shown as an example. (d) A graphical view of the SNV location in the genome browser. iJGVD, integrative Japanese Genome Variation Database; dbSNP, database single nucleotide polymorphism; SNV, single nucleotide variant.

Table 1. Number of SNVs in iJGVD by frequency class and functional category

| Functional category       | Frequency class |
|---------------------------|-----------------|
|                           | 0.05–0.10       | 0.10–0.15 | 0.15–0.20 | 0.20–0.25 | 0.25–0.30 | 0.30–0.35 | 0.35–0.40 | 0.40–0.45 | 0.45–0.50 |
| Nonsynonymous             | 3,114           | 2,113     | 1,726     | 1,393     | 1,248     | 1,170     | 995       | 1,206     | 1,268     |
| Synonymous                | 3,228           | 2,169     | 1,817     | 1,565     | 1,450     | 1,458     | 1,333     | 1,266     | 1,268     |
| 5’ UTR                    | 1,980           | 1,310     | 1,208     | 939       | 866       | 849       | 856       | 831       | 745       |
| 3’ UTR                    | 7,215           | 4,958     | 4,135     | 3,555     | 3,128     | 3,185     | 2,923     | 2,948     | 2,906     |
| Splice donor site         | 25              | 10        | 6         | 4         | 5         | 9         | 7         | 8         | 6         |
| Splice acceptor site      | 8               | 11        | 7         | 3         | 5         | 5         | 5         | 6         | 8         |
| Intron                    | 307,422         | 219,990   | 187,246   | 163,319   | 152,763   | 143,780   | 136,719   | 131,543   | 129,083   |
| Others                    | 499,044         | 366,535   | 313,854   | 283,193   | 255,771   | 245,457   | 234,201   | 229,951   | 225,074   |
| Total                     | 822,036         | 597,096   | 509,999   | 453,979   | 415,234   | 395,924   | 377,214   | 367,642   | 360,085   |

Abbreviations: iJGVD, integrative Japanese Genome Variation Database; SNVs, single nucleotide variants; UTR, untranslated region.
Allele frequency distribution for the SNVs in iJGVD was examined (Table 1). The SNV counts for each frequency class were not uniform, and the sample was enriched for low-frequency SNVs.

We compared the allele frequencies of SNVs in iJGVD with those of SNVs in HapMap3 JPT (Japanese from Tokyo) individuals (for 1,061,165 autosomal SNVs). The allele frequencies in the two populations were very similar (the correlation coefficient was 0.99). We also tested statistical difference in allele counts between ToMMo 1KJPN and HapMap JPT, and found that only a small fraction (0.022%, 226 out of 1,020,909) of SNVs showed \( P \) values of \( < 10^{-8} \) (see Supplementary Figure 2 for QQ-plots). This fraction of SNVs with small \( P \) values was very similar with that for the comparison between NGS data and SNP array data in the JPT population (Figure 2b).

SNVs in iJGVD can be searched by specifying the gene symbol, rsSNP ID, or genomic position (Figures 1b and c). Hits are displayed in a table of SNVs with allele frequencies in sequential order based on their genomic coordinates. The table can be downloaded as a text file by clicking 'Download Table.' SNVs can also be queried using the genome browser by specifying the chromosome and genomic position. The genome browser (Figure 1d) provides graphical views of the genomic location of SNVs with locations of known genes and other SNVs in dbSNP.

We constructed a public database of genomic variants with allele frequencies for the Japanese population. Variant databases for the Japanese population to date have been based on targeted SNP typing or whole-exome sequencing. iJGVD is the first database of genomic variants for Japanese individuals based on high-coverage WGS. A set of variants and the corresponding frequency information from WGS would provide a comprehensive platform for finding disease-causing variants because they can be found in non-coding regions. The allele frequencies of SNVs in iJGVD and in the HapMap3 JPT population are highly correlated (Figure 2b). Furthermore, our database contains allele frequencies for more than three million additional high-quality SNVs that were not genotyped in the HapMap3 project. We recently designed a genotyping chip, 'Japonica Array', which was optimized for the Japanese population, and probes for autosomal SNPs on Japonica Array can be seen in iJGVD.

We plan to improve the usefulness of iJGVD by adding biological annotations for SNVs and expanding search options using these annotations. Furthermore, information of linkage disequilibrium will be considered for additional data. Although iJGVD contains only SNV information at present, insertions, deletions and other structural variants will be included after quality control processes are implemented. We believe that our open variant data will be useful in medical genomics, especially for comparisons of allele frequencies in iJGVD with those of the patient group for a target disease to identify disease-causing variants.

All SNV frequency data in iJGVD are available from the National Bioscience Database Center Human Database (http://humanbdb Biosciencedbc.jp/) under accession hum0015.

ACKNOWLEDGEMENTS
This work was supported (in part) by the Tohoku Medical Megabank Project (Special Account for Reconstruction from the Great East Japan Earthquake). This research is (partially) supported by the Center of Innovation Program from Japan Science and Technology Agency, JST. All computational resources were provided by the ToMMo supercomputer system. We are indebted to all volunteers who participated in this ToMMo project. We would like to acknowledge all members associated with this project; the member list is available at the following web site: http://www.megabank.tohoku.ac.jp/english/a141201/.

COMPETING INTERESTS
The authors declare no conflict of interest.

REFERENCES
1 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
2 Halushka MK, Fan JB, Bentley K, Hisle L, Shen N, Weder A et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999; 22: 239–247.
3 Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 1999; 9: 677–679.
4 Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005; 307: 1072–1079.
5 Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. J Hum Genet 2002; 47: 605–610.

6 Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 2002; 30: 158–162.

7 Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 2002; 32: 650–654.

8 International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437: 1299–1320.

9 International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449: 851–861.

10 International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467: 52–58.

11 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56–65.

12 Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 2015; 6: 8018.

13 Katsuoka F, Yokozawa J, Tsuda K, Ito S, Pan X, Nagasaki M et al. An efficient quantitation method of next-generation sequencing libraries by using MiSeq sequencer. Anal Biochem 2014; 466: 27–29.

14 Sato Y, Kojima K, Nariai N, Yamaguchi-Kabata Y, Kawai Y, Takahashi M et al. SUGAR: graphical user interface-based data refiner for high-throughput DNA sequencing. BMC Genom 2014; 15: 664.

15 Kojima K, Nariai N, Mimori T, Yamaguchi-Kabata Y, Sato Y, Kawai Y et al. HapMonster: a statistically unified approach for variant calling and haplotyping based on phase-informative reads. Lect Notes Comput Sci 2014; 8542: 107–118.

16 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012; 6: 80–92.

17 Narahara M, Higasa K, Nakamura S, Tabara Y, Kawaguchi T, Ishii M et al. Large-scale East-Asian eQTL mapping reveals novel candidate genes for LD mapping and the genomic landscape of transcriptional effects of sequence variants. PLoS ONE 2014; 9: e100924.

18 Kawai Y, Mimori T, Kojima K, Nariai N, Danjoh I, Saito R et al. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. J Hum Genet 2015; 60: 581–587.

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