Fig. 2. (Color online) Analytic results of a case where Saffold virus was detected in a pharyngeal specimen. DNA and RNA were extracted from the specimen and sequenced. (A) Taxonomic view of MEGAN software. Saffold virus was detected in the RNA-seq reads. In this case, 0.4% of the reads were derived from the virus, which was sufficient to detect the virus in this patient. (B) Reads mapped to the reference genome of Saffold virus. The horizontal axis represents the position on the viral genome, and the vertical axis represents the abundance of mapped reads. The reads cover the whole 8 kb genome. Colored bars indicate the nucleotides where the patient's strain and the reference strain differ. The plot was made using Genome Jack software (http://www.mss.co.jp/businessfield/bioinformatics/solution/products/genomejack/).

tion over all reads, a metagenome browser can be used, such as MEGAN (11), which is one useful free software program (Fig. 2). The existence and quantity of pathogenic organisms and virulence genes are inferable from the number of detected reads, which is proportional to the number of the corresponding nucleotide sequences in the original specimen.

We developed a simple metagenomic analysis pipeline for removing ambiguous and host-derived short reads and rapidly identifying disease-causing pathogens in hospitalized patients and during outbreaks. The MePIC pipeline has a webpage interface that can be used easily by clinicians and epidemiologists, who do not have bioinformatic skill. The locally required equipment includes a benchtop next-generation DNA sequencer and a desktop PC for viewing the results. The adoption of cloud computing for metagenomic pathogen identification was proposed in the PathSeq software (12), which required bioinformatic skill for cloud computing. The MePIC pipeline, in contrast, manages the computational aspects in the background. The sequence similarity search of the database is the most computationally demanding step of metagenomic studies, and one solu-

tion is to thin the database. Using such an approach, the MetaPhlAn system (13) can speedily identify bacterial and archaeal organisms. However, we opted to maintain the entire database and rely on augmenting the computational power to hasten the analysis because clinical applications require finer taxonomic distinction: for example, the distinction of eneterohemorrhagic and commensal E. coli is critical. Within the broad possible applications of metagenomics, our pipeline is tailored for clinical use.

Metagenomic pathogen identification using next-generation sequencers surpasses conventional detection systems in sensitivity. The approach of directly sequencing nucleotides of a specimen is particularly powerful for unculturable or slow-growth pathogens (e.g., Mycobacterium). Whereas conventional PCR-based detection can miss new variants of a known pathogen due to mismatches of pre-designed primer sets, the de novo DNA/RNA sequencing approach overcomes this limitation. Metagenomic analysis can also identify a causal agent that was not known to be pathogenic (7). As for quantitative sensitivity, the metagenomic approach has been shown to be comparable to RT-PCR in

virus detection (2).

The major drawback of metagenomic pathogen identification is the cost of next-generation sequencers and reagents. Although the sequencers remain expensive, their versatile clinical and research utility (not restricted to infectious diseases) is pushing their widespread implementation in research institutes and hospitals. The rapidly decreasing reagent cost has reached approximately $100 for one million reads, which would be appropriate for pathogen identification.

The current methodology of metagenomic pathogen identification is based on sequence matches with known pathogenic species/strains. To enable detection of *unknown* pathogens, an abundant dataset of "disease cases" and "normal flora controls" is necessary. If the number of reads of an organism (which is proportional to the amount of its DNA in the specimen) is much larger in cases than controls, infection by the organism can be suspected. The development of such pathogen discovery will require the accumulation of metagenomic data for disease-causing and normal flora and the invention of analytic tools. We believe that the use of the MePIC pipeline will promote metagenomic pathogen identification and improve the understanding of infectious diseases.

The source code for installing on a local server is available from the authors upon request. The website of the pipeline is https://mepic.nih.go.jp/. The sequence reads of the pharyngeal specimen that included the Saffold virus are available from the DDBJ Sequence Read Archive under accession number DRA000973.

REFERENCES

1. Tang, P. and Chiu, C. (2010): Metagenomics for the discovery of novel human viruses. Future Microbiol., 5, 177–189.
2. Chan, J.Z-M., Pallen, M.J., Oppenheim, B., et al. (2012): Genome sequencing in clinical microbiology. Nat. Biotechnol., 30, 1068–1071.
3. Kuroda, M., Sekizuka, T., Shinya, F., et al. (2012): Detection of a possible bioterrorism agent, *Francisella* sp., in a clinical specimen using next-generation direct DNA sequencing. J. Clin. Microbiol., 50, 1810–1812.
4. Katano, H., Sato, S., Sekizuka, T., et al. (2012): Pathogenic characterization of a cervical lymph node derived from a patient with Kawasaki disease. Int. J. Clin. Exp. Pathol., 5, 814–823.
5. Kuroda, M., Katano, H., Nakajima, N., et al. (2010): Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by *de novo* sequencing using a next-generation DNA sequencer. PLoS ONE, 5, e10256.
6. Mizuta, K., Kuroda, M., Kurimura, M., et al. (2012): Epidemic myalgia in adults associated with human parechovirus type 3 infection, Yamagata, Japan, 2008. Emerg. Infect. Dis., 18, 1787–1793.
7. Kawai, T., Sekizuka, T., Yahata, Y., et al. (2012): Identification of *Kudoa septempunctata* as the causative agent of novel food poisoning outbreaks in Japan by consumption of *Paralichthys olivaceus* in raw fish. Clin. Infect. Dis., 54, 1046–1052.
8. Li, H. and Durbin, R. (2009): Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.
9. Morgulis, A., Coulouris, G., Raytselis, Y., et al. (2008): Database indexing for production MegaBLAST searches. Bioinformatics, 24, 1757–1764.
10. Altschul, S.F., Gish, W., Miller, W., et al. (1990): Basic local alignment search tool. J. Mol. Biol., 215, 403–410.
11. Huson, D.H., Mitra, S., Ruscheweyh, H-J., et al. (2011): Integrative analysis of environmental sequences using MEGAN4. Genome Res., 21, 1552–1560.
12. Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., et al. (2011): PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol., 29, 393–396.
13. Segata, N., Waldron, L., Ballarini, A., et al. (2012): Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods, 9, 811–814.