Jpn. J. Infect. Dis., 67, 62-65, 2014

Short Communication

MePIC, Metagenomic Pathogen Identification for Clinical Specimens

Fumihiko Takeuchi[1], Tsuyoshi Sekizuka[1], Akifumi Yamashita[1], Yumiko Ogasawara[1], Katsumi Mizuta[2], and Makoto Kuroda[1]*

[1]*Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo 162-8640; and* [2]*Department of Microbiology, Yamagata Prefectural Institute of Public Health, Yamagata 990-0031, Japan*

(Received July 10, 2013. Accepted September 26, 2013)

SUMMARY: Next-generation DNA sequencing technologies have led to a new method of identifying the causative agents of infectious diseases. The analysis comprises three steps. First, DNA/RNA is extracted and extensively sequenced from a specimen that includes the pathogen, human tissue and commensal microorganisms. Second, the sequenced reads are matched with a database of known sequences, and the organisms from which the individual reads were derived are inferred. Last, the percentages of the organisms' genomic sequences in the specimen (i.e., the metagenome) are estimated, and the pathogen is identified. The first and last steps have become easy due to the development of benchtop sequencers and metagenomic software. To facilitate the middle step, which requires computational resources and skill, we developed a cloud-computing pipeline, MePIC: ''Metagenomic Pathogen Identification for Clinical specimens.'' In the pipeline, unnecessary bases are trimmed off the reads, and human reads are removed. For the remaining reads, similar sequences are searched in the database of known nucleotide sequences. The search is drastically sped up by using a cloud-computing system. The webpage interface can be used easily by clinicians and epidemiologists. We believe that the use of the MePIC pipeline will promote metagenomic pathogen identification and improve the understanding of infectious diseases.

Next-generation DNA sequencing technologies have led to a new method of identifying the causative agent of infectious diseases in hospitalized patients and during outbreaks (1,2). By directly sequencing millions of DNA/RNA molecules in a specimen and matching the sequences to those in a database, pathogens can be inferred. The analysis comprises three steps. First, the nucleotide sequences of the specimen, which includes the pathogen, human tissue and commensal microorganisms, are read using a next-generation sequencer. Second, from bioinformatic processing of the reads, the organisms from which the individual reads were derived are inferred. Last, the percentages of the organisms' genomic sequences in the specimen are estimated, and the pathogen is identified. Although the first and last steps have become easy due to the development of benchtop sequencers and metagenomic software, the middle step still requires computational resources and bioinformatic skill. To facilitate the middle step, we developed a cloud-computing pipeline that is easy and fast.

The prototype of the pipeline has been used in our metagenomic search for pathogens in various clinical cases. We have reported metagenomic analyses of clinical specimens that successfully identified *Francisella tularensis* in an abscess as a pathogen (3), *Streptococcus* spp. in a lymph node as a possible causative candidate of Kawasaki disease (4) and heterogeneity of the 2009 pandemic influenza A virus (A/H1N1/2009) in the lung (5). In an outbreak of 22 adults with myalgia, the majority were infected with human parechovirus type 3, which typically causes disease in young children (6). In recent food poisoning outbreaks that were due to raw fish consumption, a flounder parasite *Kudoa septempunctata* was discovered as the causative agent (7). In all cases, the pathogen identification was primarily due to metagenomic analyses using next-generation sequencers.

In the workflow of metagenomic pathogen identification using MePIC, the first step is performed by the user. DNA and/or RNA is extracted from a specimen, such as sputum, feces, an abscess or blood, and a library of DNA/cDNA is prepared for sequencing. The library is sequenced using a benchtop next-generation sequencer, and the sequenced reads are uploaded to the MePIC pipeline via a secure internet connection. The pipeline accepts input files in FASTQ format, which is the standard for next-generation sequencing analysis. When using Illumina MiSeq sequencers (San Diego, Calif., USA).

In the second step, the uploaded reads are processed by the MePIC pipeline (Fig. 1). Unnecessary adapter sequences and low quality bases are trimmed off the reads using the fastq-mcf program in the ea-utils package (http://code.google.com/p/ea-utils/). Human-derived reads are detected through comparisons with the human genome using the BWA (8) program; the

* Corresponding author: Mailing address: Pathogen Genomics Center, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan. Tel: +81-3-5285-1111, Fax: +81-3-5285-1166, E-mail: makokuro@niid.go.jp

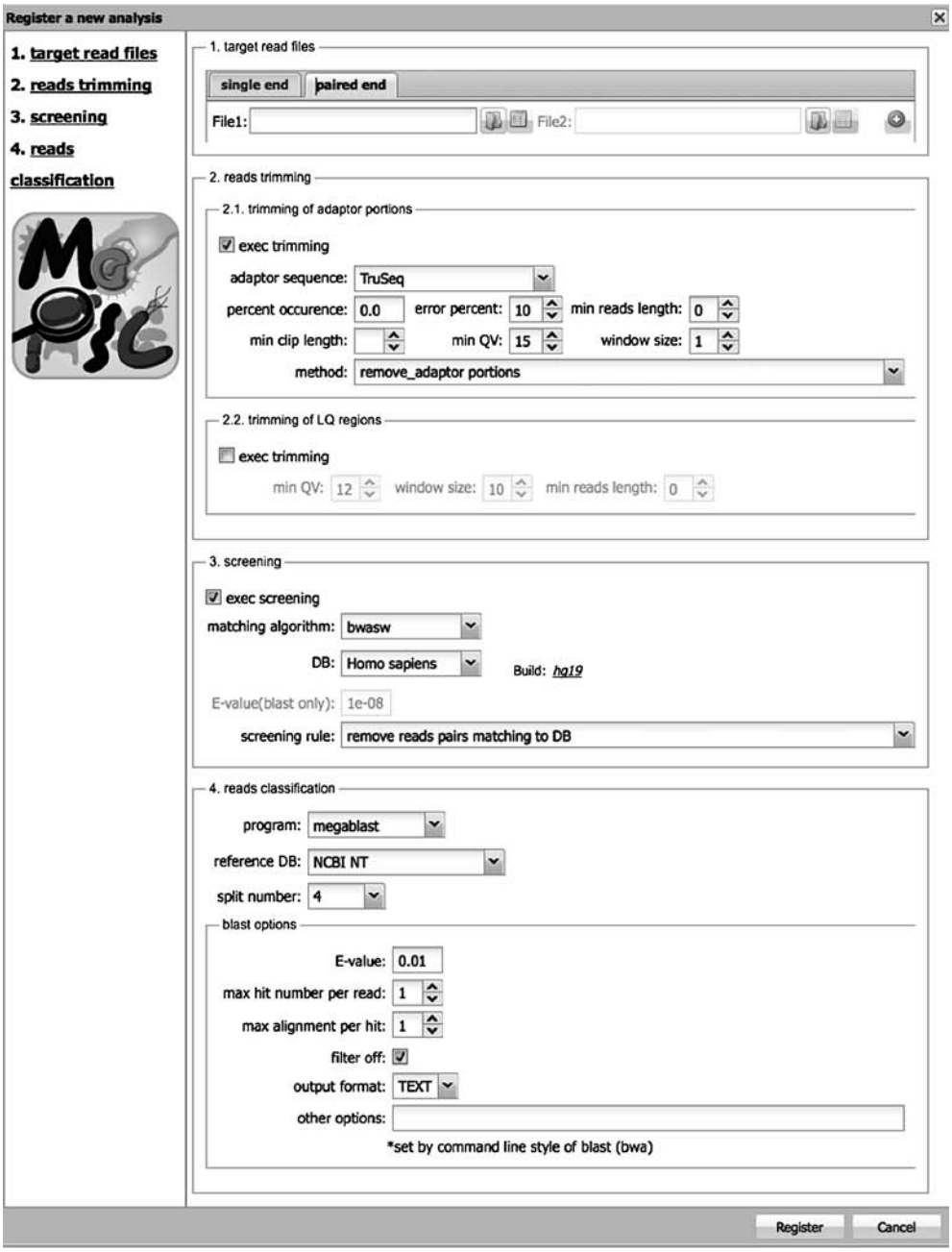

Fig. 1. Screenshot of the MePIC pipeline. In box 1, the user specifies the next-generation sequencer reads for upload. In box 2, details are set for trimming adaptor sequences and low quality bases from the reads. In box 3, criteria are set for the exclusion of human reads. In box 4, the user chooses the program for searching the database of known sequences.

number of human reads are counted, but the reads themselves are removed from the downstream analysis. For each of the remaining reads, similar sequences are searched in the database of all known nucleotide sequences (NCBI nt) using the MEGABLAST (9) or BWA program (10). Based on the information of the database sequences that match with the read, we can infer the gene (e.g., virulence gene) and organism (e.g., *Escherichia coli*) that the read is derived from.

The run time of the pipeline is primarily allotted to searching the database of known sequences. The required time can be drastically shortened by splitting the job and running in parallel using a cloud-computing system or local server. Respectively, it takes 10 h for one core of 2.67 GHz and 6 min for 100 cores to perform MEGABLAST search against the nt nucleotide database (as of year 2013) for one million reads of length 200 bp. The run time varies according to the sample source and condition. In blood samples, $>90\%$ of reads are derived from human and removed in the preprocessing step, and accordingly the time for database search of the remaining reads is reduced. Human derived reads are less in sputum samples (60%) or normal feces ($\sim 0\%$), which accordingly demands more database search time.

In the final step of the workflow, the user downloads the database search result to a local PC, including the reads annotated with the organism and gene function. To summarize the taxonomic and functional informa-

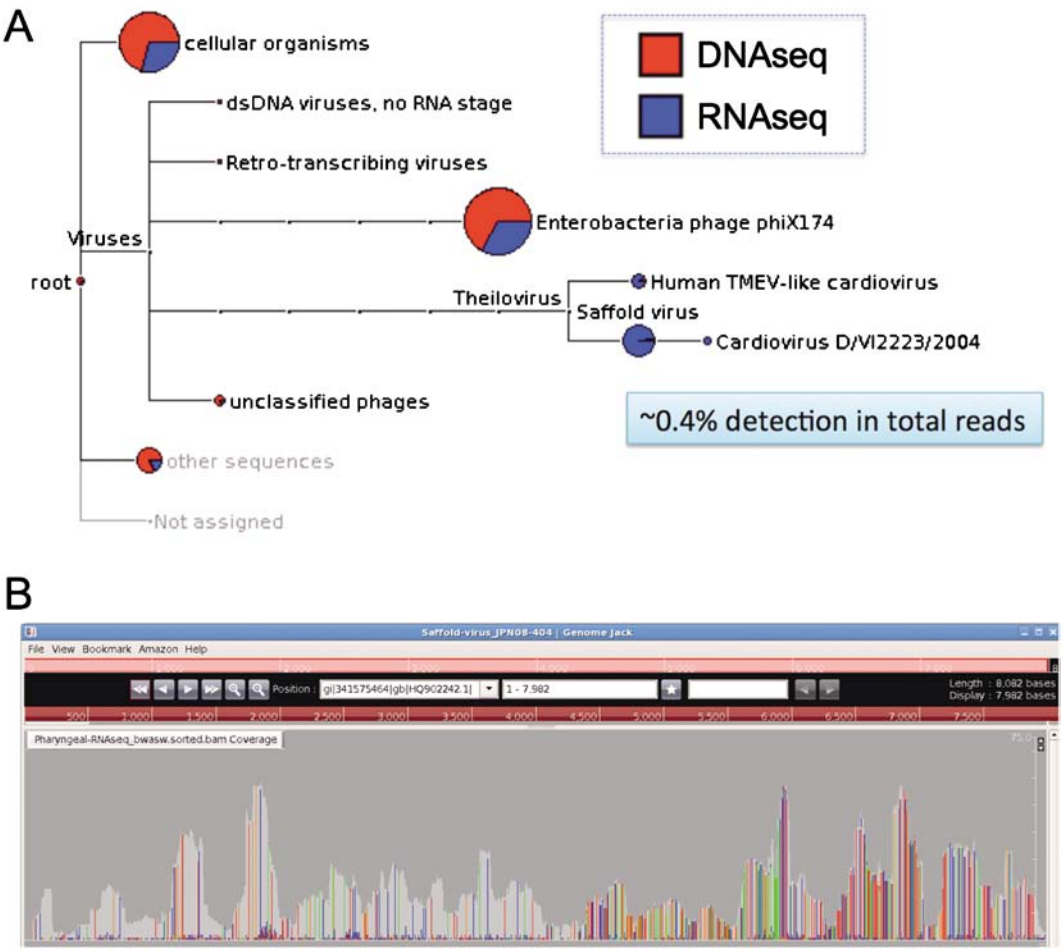

Fig. 2. (Color online) Analytic results of a case where Saffold virus was detected in a pharyngeal specimen. DNA and RNA were extracted from the specimen and sequenced. (A) Taxonomic view of MEGAN software. Saffold virus was detected in the RNA-seq reads. In this case, 0.4% of the reads were derived 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 solution 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.

Acknowledgments This work was supported by a grant for Research on Emerging and Re-emerging Infectious Diseases (H23 Shinko-Ippan-007, H25 Shinko-Ippan-015) from the Ministry of Health, Labour and Welfare, Japan, and was also supported by a grant-in-aid for Scientific Research (C) from the Japan Society for the Promotion of Science (Grant No. 23590527). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The development of MePIC pipeline was partially supported by Illumina Inc.

Conflict of interest None to declare.

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.