frontiers in Microbiology

PERSPECTIVE
published: 26 July 2016
doi: 10.3389/fmicb.2016.01061

Analysis of Metagenomics Next Generation Sequence Data for Fungal ITS Barcoding: Do You Need Advance Bioinformatics Experience?

*Abdalla Ahmed**

College of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia

During the last few decades, most of microbiology laboratories have become familiar in analyzing Sanger sequence data for ITS barcoding. However, with the availability of next-generation sequencing platforms in many centers, it has become important for medical mycologists to know how to make sense of the massive sequence data generated by these new sequencing technologies. In many reference laboratories, the analysis of such data is not a big deal, since suitable IT infrastructure and well-trained bioinformatics scientists are always available. However, in small research laboratories and clinical microbiology laboratories the availability of such resources are always lacking. In this report, simple and user-friendly bioinformatics work-flow is suggested for fast and reproducible ITS barcoding of fungi.

Keywords: ITS sequencing, next generation sequencing, fungi, bioinformatic tools and databases, metagenomics

OPEN ACCESS

Edited by:
Vijai Kumar Gupta,
NUI Galway, Ireland

Reviewed by:
Jeanette Wagener,
University of Aberdeen, UK
Venkataramana M.,
DRDO-Bharathiar University Center for Life Sciences, India

***Correspondence:**
Abdalla Ahmed
aoaahmed@uqu.edu.sa

Specialty section:
This article was submitted to Fungi and Their Interactions, a section of the journal Frontiers in Microbiology

Received: *12 February 2016*
Accepted: *23 June 2016*
Published: *26 July 2016*

Citation:
Ahmed A (2016) Analysis of Metagenomics Next Generation Sequence Data for Fungal ITS Barcoding: Do You Need Advance Bioinformatics Experience? Front. Microbiol. 7:1061. doi: 10.3389/fmicb.2016.01061

INTRODUCTION

Since the introduction of Sanger Sequencing, many microbiology laboratories started using DNA sequence data for microbial identification and genotyping. These DNA sequence data revolutionized microbial genotyping and taxonomy and quickly became part of the routine clinical microbiology work (Makimura, 2001; Leaw et al., 2006). DNA sequence data generated by Sanger Sequencing technology characterized by relatively limited size (±800 bases single read) and high base calling quality. This nature of Sanger sequence data enable most scientists, with no standard bioinformatics training, to perform many basic sequence data analysis without the need of highly trained bioinformatics specialists. However, with the introduction of next-generation sequencing (NGS) technologies, huge sequence data with varying degrees of quality become available. The analysis of such large and complex data become rather difficult. Therefore, it become mandatory, for many research centers, to recruit specially trained bioinformatics staff to handle the huge NGS data obtained from these diverse sequencing platforms. Alternatively, many small research centers and microbiology laboratories are forced to seek help in data analysis from specialized sequencing centers or bioinformatics commercial services providers.

In recent years, and with the great development of NGS platforms and sequencing technologies, DNA sequencing in no longer done in specialized sequencing centers and reference research laboratories only. Library preparation protocols for NGS become simple and acquisition of next generation sequencers become affordable by many research and diagnostics laboratories. Therefore, it become important for microbiologists to know how to make sense of the massive NGS data generated by these new sequencing technologies. In reference laboratories, the analysis

of such data is not a big deal, since suitable IT infrastructure and well-trained bioinformaticians are always available. However, in small research laboratories and clinical microbiology laboratories the availability of such resources are always lacking.

Microbial DNA sequencing applications are numerous and these applications are rapidly evolving with introduction new sequencing technologies. In clinical microbiology, NGS data can be used in many routine applications. For example, whole microbial genome sequencing and targeted sequencing are currently widely used for unlimited applications such as species identification, virulence genes detection, antimicrobial resistant mechanisms prediction and genotyping (Zankari et al., 2012; Joensen et al., 2014; Larsen et al., 2014; Garnaud et al., 2015). Another interesting area for microbiologist is metagenomics, which can be used for sequencing of novel species from environmental specimens. Metagenomics can also be used for species identification of bacteria and fungi by targeted sequencing of the 16S and ITS regions of the rRNA genes, respectively (Salipante et al., 2013; Tang et al., 2015).

WHY DO WE NEED TO SEQUENCE DNA FOR SPECIES IDENTIFICATION?

Species identification in fungi is difficult and time-consuming even for those with special training and experience in medical mycology. Therefore, it becomes routine in many centers to sequence the Internal Transcribed Spacer (ITS) region of the ribosomal RNA genes (rDNA) for species identification. The rDNA of fungi exist as a multiple-copy gene family comprised of highly similar DNA sequences (typically from 8 to 12 kb each). The ITS region of the rDNA is the most widely sequenced DNA region in fungi. ITS is typically most useful for molecular systematics at the species level, and even within species. This is because ITS characterized by high degree of variation than other regions of rDNA such as small sub unit (SSU) and large sub unit (LSU) of the rDNA.

WHY DO WE NEED HELP IN "BIOINFORMATICS" BUT NOT IN "DNA SEQUENCING"?

Sequencing library preparation workflow is getting much easier. Thanks for the innovative, simple and quick library preparation protocols for DNA sequencing. However, data analysis remains the most challenging step in this wonderful technology. NGS data analysis is the biggest challenge in routine application of NGS in clinical setting (Desai and Jere, 2012). NGS data analysis is rapidly evolving field, but still largely carried out using commercial and/or open source research tools not designed for clinical laboratories (Desai and Jere, 2012). One another issue on NGS data analysis is the huge amount of data generated, which is beyond the computing infrastructure of most clinical setting (Stein, 2010). In fact, most of NGS platforms have some data analysis functionality, which can be done on the same sequencing machine. However, these automated bioinformatics workflows does not provide total analysis solutions, and it remains difficult and unclear for many microbiologist how to quickly and efficiently analyze the raw sequencing data to get clear answers for many basic questions.

In this report, we present a simple and easy bioinformatics workflow for one of the commonly asked questions, "what is the species of this fungal isolate"? The workflow consist of sequences quality check, *de novo* assembly and sequence similarity search. The workflow is based on two genomics computing environments Illumina BaseSpace[1] and the Public Galaxy Server (Galaxy Project[2]). This bioinformatics workflow needs only basic bioinformatics knowledge, and can be done by any scientist using any computer connected to the internet.

SIMPLE DATA ANALYSIS WORKFLOW:

Regardless of NGS platform used, sequence data normally stored in text file in a Fastq format, which contains sequence data and the quality score of base calling for each base. This Fastq file is your starting material. If you are using paired end library you will end with two Fastq files one for each read (read 1 and read 2). Before start analyzing the data, raw sequence reads need to checked for quality. The most important quality parameters are quality score of base calling, number of reads and reads length distribution. In addition, sequence reads need to be checked for possible sequencing adapter contamination, especially if using small sequencing target. Low quality sequence reads and/or sequence contamination need to be removed from data sets before any subsequent data manipulation or analysis. For quality check of sequencing read, we recommend the use of FastQC tool, which is available as a push-button tool at the Public Galaxy Server[2]) and Illumina BaseSpce (BaseSpace Labs, Illumina, San Diego, CA, USA).

Once sequence data has been checked for quality, the next step is to assemble the sequence reads into contigs using any short sequence *de novo* assembler. The aim of this *de novo* assembly is to covert the large number of reads into few contigs (a set of overlapping DNA segments that together represent a consensus region of DNA). Assembled contigs can be easily used for sequence similarity search and species identification. In this workflow, we recommend the use of Velvet assembler or SPAdes Genome Assembler 3.0 for the *de novo* assembly. These two applications are in the Illumina BaseSpace applications[1]. Sequence reads can also be assembled using many other free or commercial tools. Once assembly is finished, assembled contigs can visualized using any text viewer such as Notepad or the Universal Viewer[3]. The best contigs with sizes matching the expected sequences regions can be directly used for ITS based

[1]https://basespace.illumina.com

[2]https://usegalaxy.org

[3]http://www.uvviewsoft.com/uviewer/

species identification at the NCBI Nucleotide BLAST[4] or the ISHAM ITS database[5] (Irinyi et al., 2015).

CONCLUSION

Analysis of NGS data for ITS-based fungal identification is easy to perform and does not require advance bioinformatics training or expensive IT infrastructure. However, in addition to the available bioinformatics tools, there is a need for more automated data interpretation tools, which are able to generate easily understandable clinical reports. When such tools become available, NGS-based identification and other microbial sequence-based applications will be part of the clinical microbiology routine work.

[4] http://blast.ncbi.nlm.nih.gov/Blast.cgi

[5] http://its.mycologylab.org/Biolomicssequences.aspx

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

King Abdul-Aziz City for Science and Technology, Riyadh, Saudi Arabia. NSTIP grant number 12-BIO2 295-10.

REFERENCES

Desai, A. N., and Jere, A. (2012). Next-generation sequencing: ready for the clinics? *Clin. Genet.* 81, 503–510. doi: 10.1111/j.1399-0004.2012.01865.x

Garnaud, C., Botterel, F., Sertour, N., Bougnoux, M. E., Dannaoui, E., Larrat, S., et al. (2015). Next-generation sequencing offers new insights into the resistance of Candida spp. to echinocandins and azoles. *J. Antimicrob. Chemother.* 70, 2556–2565. doi: 10.1093/jac/dkv139

Irinyi, L., Serena, C., Garcia-Hermoso, D., Arabatzis, M., Desnos-Ollivier, M., Vu, D., et al. (2015). International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database–the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med. Mycol.* 53, 313–337. doi: 10.1093/mmy/myv008

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., et al. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 52, 1501–1510. doi: 10.1128/JCM.03617-13

Larsen, M. V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., et al. (2014). Benchmarking of methods for genomic taxonomy. *J. Clin. Microbiol.* 52, 1529–1539. doi: 10.1128/JCM.02981-13

Leaw, S. N., Chang, H. C., Sun, H. F., Barton, R., Bouchara, J. P., and Chang, T. C. (2006). Identification of medically important yeast species by sequence analysis of the internal transcribed spacer regions. *J. Clin. Microbiol.* 44, 693–699. doi: 10.1128/JCM.44.3.693-699.2006

Makimura, K. (2001). Species identification system for dermatophytes based on the DNA sequences of nuclear ribosomal internal transcribed spacer 1. *Nippon Ishinkin Gakkai Zasshi* 42, 61–67. doi: 10.3314/jjmm.42.61

Salipante, S. J., Sengupta, D. J., Rosenthal, C., Costa, G., Spangler, J., Sims, E. H., et al. (2013). Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS ONE* 8:e65226. doi: 10.1371/journal.pone.0065226

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol.* 11, 207. doi: 10.1186/gb-2010-11-5-207

Tang, C. Y., Yiu, S. M., Kuo, H. Y., Tan, T. S., Liao, K. H., Liu, C. C., et al. (2015). Application of 16S rRNA metagenomics to analyze bacterial communities at a respiratory care centre in Taiwan. *Appl. Microbiol. Biotechnol.* 99, 2871–2881. doi: 10.1007/s00253-014-6176-7

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi: 10.1093/jac/dks261

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.