

Metagenomic approach for discovering new pathogens in infection disease outbreaks

Emanuela Giombini, Barbara Bartolini, Gabriella Rozera, Marina Selleri, Paola Zaccaro, Maria Rosaria Capobianchi, Isabella Abbate

Laboratory of Virology, National Institute of Infectious Diseases "L. Spallanzani", Rome, Italy

Key words: Metagenomics, Viroma, Next generation sequencing, Outbreak

SUMMARY

Viruses represent the most abundant biological components on earth. They can be found in every environment, from deep layers of oceans to animal bodies. Although several viruses have been isolated and sequenced, in each environment there are millions of different types of viruses that have not been identified yet. The advent of next-generation sequencing technologies with their high throughput capabilities make possible to study in a single experiment all the community of microorganisms present in a particular sample "microbioma". They made more feasible the application of the metagenomic approach, by which it is also possible to discover and identify new pathogens, that may pose a threat to public health. This paper summarizes the most recent applications of next-generation sequencing to discover new viral pathogens during the occurrence of infection disease outbreaks.

Metagenomic is an experimental methodology that combines molecular biology and genomics to identify the complex mixture of organisms contained in a given environment. This approach has two steps: first massive sequencing of all the genomic material present in a given environment; second the identification of each organism present by searching in genome databases (Figure 1).

In particular, the process termed "viral metagenomics or viroma study" focalizes the attention only on virus organisms. Viruses can be found in a wide range of environments, from extreme sea environments such as deep layers of the oceans (9), food (11) potable water (12, 15), soil (7), and even in several microenvironments contained in human bodies (2, 3, 5, 17, 18).

An important feature in studying viruses is represented by the intracellular nature of these organisms. One limitation is the impossibility to grow a number of them under standard laboratory conditions. This feature has drastically reduced in the past the capability of identification of new viruses (4).

In recent days, the high throughput of next generation sequencing techniques is having a great impact in expanding knowledge on viruses, especially those displaying a very high genomic diversity, such as HIV and HBV (13, 14). Together with HCV and others, these viruses are present in an infected host as a swarm of highly genetically related variants, referred to as quasispecies. Next generation sequencing allow to study directly on the clinical specimen the

genetic diversity of the viral quasispecies, avoiding biases that may be introduced by *in vitro* replication.

Today, the most widely used next generation sequencing platforms are three: the Genome Sequencer from Roche 454 Life Sciences (www.454.com), the Solexa Genome Analyzer from Illumina (www.illumina.com), the SOLiD System from Applied Biosystems (www.appliedbiosystems.com).

All of them have an high throughput ranging, from hundreds of thousands of sequences (reads) per run of 454 technology, to millions of reads usually obtained by the application of the other two technologies. At the moment, 454 technology gives the longest reads approximately up to 700-800 nucleotides (nt), while the others Solexa/Illumina and ABI/SOLiD, can generate, in average, reads of only 35-50 nt. All of these next generation sequencing approaches may render possible to analyze a given microbial community in a single run.

By metagenomic approach, it is possible to discover and identify new pathogens, that may constitute a threat for public health.

One of the first examples of the use of this approach to discover new pathogens was the identification of a new Arenavirus in a cluster of fatal transplant-associated diseases (10). RNA samples from the tissues of the victims were sequenced using ABI/SOLiD method, obtaining, in few hours, hundred thousands of sequences among which few of them related to Old World

Corresponding author: Isabella Abbate

Department of Virology, National Institute of Infectious Diseases "L. Spallanzani"
Via Portuense, 292 - Rome, Italy - Tel:+39 06 55170655; Fax:+39 06 5594555
E-mail: isabella.abbate@inmi.it

Arenavirus group. Short after, specific PCR and phylogenetic analysis classified as belonging to a new member of the *Arenaviridae* family the strains sequenced by next generation approach. The same group of researchers also found by metagenomic approach a new Arenavirus in an outbreak of five cases of hemorrhagic fever, four of them fatal, occurred in South of Africa in September 2008 (6). Using ABI/SOLiD method, they obtained between 87,500 and 106,500 sequence reads from RNA extracts from liver biopsies and serum. In only 72 hours, it was possible to sequence approximately 50% of the genome, which was identified as a new member of the *Arenaviridae* family.

Another example of the utility of the metagenomic approach for public health was the rapid identification and characterization of the new pandemic influenza A virus (H1N1) in 2009 (8).

As in other biomedical reference centres, in our laboratory the high throughput sequencing approach was applied to study this new pandemic H1N1 virus in samples from infected patients. By 454 technology, we were able, directly in nasopharyngeal swabs, to obtain enough reads to reconstruct the entire viral genome in a single run. This allowed us to study the viral quasispecies without the biases introduced by in vitro replication of the virus that inevitably adds additional artificial diversity (1).

The application of viral metagenomics can also be used to better characterize infectious diseases with uncertain aetiology. As an example, 454 technology

was able to identify, with respect to standard sequencing, a larger number of new species of viruses in stools samples from Asian children affected by non polio acute flaccid paralysis (AFP), expanding the possibility to find the real aetiological agent of this orphan disease (16).

The metagenomic approach could be useful to reveal potential pathogens in substances included in human diets. In 2010 Park et al. found that some viruses could modify the bacterial ecosystem present in fermented food that, in turn, could alter human gastrointestinal microbiota (11).

Although the cost and the complexity of the metagenomic approach, in particular of the bioinformatic tools, that have to be implemented in parallel with advances in molecular sequencing, it is reasonable that in the future more and more data will be obtained by this methodology in different human fields enhancing our capabilities to challenge emerging diseases and new pandemics.

REFERENCES

1. Bartolini B, Chillemi G, Abbate I, et al. Assembly and Characterization of Pandemic Influenza a H1N1 Genome in Nasopharyngeal Swabs Using High-Throughput Pyrosequencing. *New Microbiol*, 2011; in press.
2. Breitbart M, Haynes M, Kelley S, et al. Viral Diversity and Dynamics in an Infant Gut. *Res Microbiol*, 2008; 159(5): 367-73.
3. Breitbart M, Hewson I, Felts B, et al. Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *J Bacteriol*, 2003; 185(20): 6220-3.

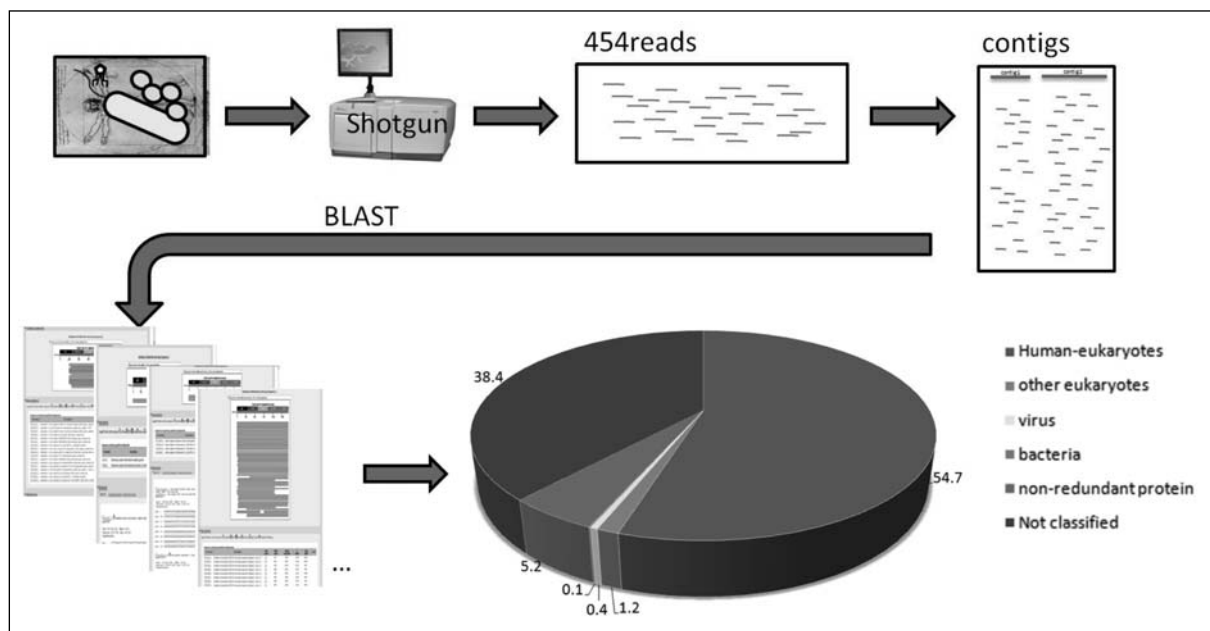


Figure 1. A schematic representation of the work flow to carry out a metagenomic study. First: the nucleic acid material is collected from the environment and undergo massive sequencing. Second: through a bioinformatic pipeline, consisting in a number of searches against different nucleotide databases, with a growing level of homology stringency, a detailed quantitative description of the different organisms present in the sample is obtained.

4. Breitbart M, Rohwer F. Here a Virus, There a Virus, Everywhere the Same Virus? *Trends Microbiol*, 2005; 13(6): 278-84.
5. Breitbart M, Rohwer F. Method for Discovering Novel DNA Viruses in Blood Using Viral Particle Selection and Shotgun Sequencing. *Biotechniques*, 2005; 39(5): 729-36.
6. Briese T, Paweska JT, McMullan LK, et al. Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever-Associated Arenavirus from Southern Africa. *PLoS Pathog*, 2009; 5(5): e1000455.
7. Fierer N, Breitbart M, Nulton J, et al. Metagenomic and Small-Subunit Rna Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. *Appl Environ Microbiol*, 2007; 73(21): 7059-66.
8. Greninger AL, Chen EC, Sittler T, et al. A Metagenomic Analysis of Pandemic Influenza a (2009 H1N1) Infection in Patients from North America. *PLoS One*, 2010; 5(10): e13381.
9. Heidelberg KB, Gilbert JA, Joint I. Marine Genomics: At the Interface of Marine Microbial Ecology and Biodiscovery. *Microb Biotechnol*, 2010; 3(5): 531-43.
10. Palacios G, Druce J, Du L, et al. A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. *N Engl J Med*, 2008; 358(10): 991-8.
11. Park EJ, Kim KH, Abell GC, et al. Metagenomic Analysis of the Viral Community in Fermented Foods. *Appl Environ Microbiol*, 2010.
12. Rosario K, Nilsson C, Lim YW, et al. Metagenomic Analysis of Viruses in Reclaimed Water. *Environ Microbiol*, 2009; 11(11): 2806-20.
13. Rozera G, Abbate I, Bruselles A, et al. Massively Parallel Pyrosequencing Highlights Minority Variants in the Hiv-1 Env Quasispecies Deriving from Lymphomonocyte Sub-Populations. *Retrovirology*, 2009; 6: 15.
14. Solmone M, Vincenti D, Prosperi MC, et al. Use of Massively Parallel Ultradeep Pyrosequencing to Characterize the Genetic Diversity of Hepatitis B Virus in Drug-Resistant and Drug-Naive Patients and to Detect Minor Variants in Reverse Transcriptase and Hepatitis B S Antigen. *J Virol*, 2009; 83(4): 1718-26.
15. Venter JC, Remington K, Heidelberg JF, et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 2004; 304(5667): 66-74.
16. Victoria JG, Kapoor A, Li L, et al. Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis. *J Virol*, 2009; 83(9): 4642-51.
17. Willner D, Furlan M, Haynes M, et al. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. *PLoS One*, 2009; 4(10): e7370.
18. Zhang T, Breitbart M, Lee WH, et al. Rna Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. *PLoS Biol*, 2006; 4(1): e3.