Rev. Sci. Tech. Off. Int. Epiz., 2016, **35** (1), 83-93

Metagenomic approaches to identifying infectious agents

D. Höper [1], T.C. Mettenleiter [2] & M. Beer [1]

(1) Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Institute of Diagnostic Virology, Südufer 10, D-17493 Greifswald – Insel Riems, Germany
(2) Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Institute of Molecular Virology and Cell Biology, Südufer 10, D-17493 Greifswald – Insel Riems, Germany
Corresponding author: Dirk.Hoeper@fli.bund.de

Summary

Since the advent of next-generation sequencing (NGS) technologies, the untargeted screening of samples from outbreaks for pathogen identification using metagenomics has become technically and economically feasible. However, various aspects need to be considered in order to exploit the full potential of NGS for virus discovery. Here, the authors summarise those aspects of the main steps that have a significant impact, from sample selection through sample handling and processing, as well as sequencing and finally data analysis, with a special emphasis on existing pitfalls.

Keywords

Analysis software – Metagenome – Next-generation sequencing – Sample handling – Sample treatment – Virus discovery.

Introduction

The first metagenomics analyses aimed to identify microbes found in diverse habitats (1), and were based on sequencing ribosomal ribonucleic acid (rRNA) in order to describe the microbial diversity as completely as possible, independent of the isolation of individual strains. Later, the metabolic functions, as encoded in the sequenced DNA or RNA fragments, were also assessed (2). Since the advent of the various next-generation sequencing (NGS) platforms, metagenomics has also had an impact on veterinary science and pathogen detection. The power of NGS-driven metagenomic approaches to identify infectious agents is defined by the vast amount of sequencing information that can now be obtained in a single experiment using novel sequencing instruments. Moreover, the sequence information gained can be useful for downstream analyses such as reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) screening of additional samples, for instance for confirmation, quantification and epidemiological purposes.

Initially, the term 'next-generation sequencing' was coined to differentiate new high-throughput sequencing methods from classical Sanger sequencing. Currently, NGS techniques are further divided into second- and third-generation sequencing methods. While the technologies of the second generation require clonal amplification of the library in order to generate sufficient signal intensity for detection, third-generation technologies directly sequence the original unamplified library (i.e. only the original input molecules). Given that the datasets obtained by third-generation sequencing are currently not large enough to provide the necessary depth for a metagenomics analysis, this paper focuses on second-generation sequencing only.

All second-generation sequencers share a number of important characteristics which define the value of NGS for pathogen detection. The DNA libraries necessary for sequencing are prepared from the input DNA using only molecular biological techniques. While the amplification of individual plasmid clones in bacterial cultures was required for high-throughput Sanger sequencing, the clonal amplification that is necessary to gain sufficient

doi: http://dx.doi.org/10.20506/rst.35.1.2419

signal strength for sequencing is now also done using only molecular techniques. The sample preparation procedures for NGS all aim to ensure an unbiased representation of the input DNA in the sequence output, i.e. the detection of all nucleic acids accessible in the sample with equal probability; however, this is not always achieved (3). Taken together, the sample preparation techniques and sequencing technologies for all NGS platforms enable sequencing of millions or even billions of individual molecules in parallel. Hence, datasets comprising a huge number of short sequence fragments 70–800 base pairs in length (so-called 'reads'), each reflecting individual input molecules, can be generated.

These sequence fragments can be analysed both qualitatively and quantitatively. Qualitative analyses mainly focus on the detection of all individual sequences present in a sample, for instance by assembly of larger sequences from the generated reads, without taking into account the abundances of reads representing certain species. The latter is also possible, however, assuming that sample preparation is unbiased. The determination of the relative abundances of different species by counting the reads representing each species is therefore feasible. This of course requires the unambiguous classification of reads to species.

The composition of the organisms and viruses present in the sample can be determined after classification, for example by comparison to sequences available in databases. In the case of a disease of unknown cause, even a single read may give the initial information necessary to identify the causative agent and start follow-up analyses, including further sequencing efforts and PCR development and testing.

Metagenomics has had a significant input into veterinary science and diagnosis in recent years, with the detection of several newly emerging or re-emerging diseases. An emerging disease of high impact which was first identified by metagenomics is caused by Schmallenberg virus (SBV) (4), which is an orthobunyavirus of the Simbu serogroup. The virus was detected, using an NGS-based approach, in blood samples from cows with milk drop and fever in November 2011; SBV and related viruses had never been detected in Europe before. Conventional diagnostic systems had failed to identify the reason for the clinical signs observed in summer and autumn 2011 in cattle in the Netherlands, Belgium and Germany (5). Further examples are the identification of astroviruses in cattle diagnosed with bovine encephalitis in Europe (6), influenza D virus in cattle in the United States and France (7, 8), bat influenza viruses H17N10 and H18N11 in bats in Central and South America (9, 10), and a novel zoonotic bornavirus in variegated squirrels in Germany (11). Virus discovery is therefore a key domain of NGS-based metagenomics, and both outbreak diagnosis and screening investigations are successful strategies.

Figure 1 schematically shows the main steps of metagenomic pathogen identification. All steps of the complete procedure may have a strong impact on the results obtained. The following paragraphs describe the specific issues for all the steps depicted in Figure 1 that need to be considered for successful pathogen detection.

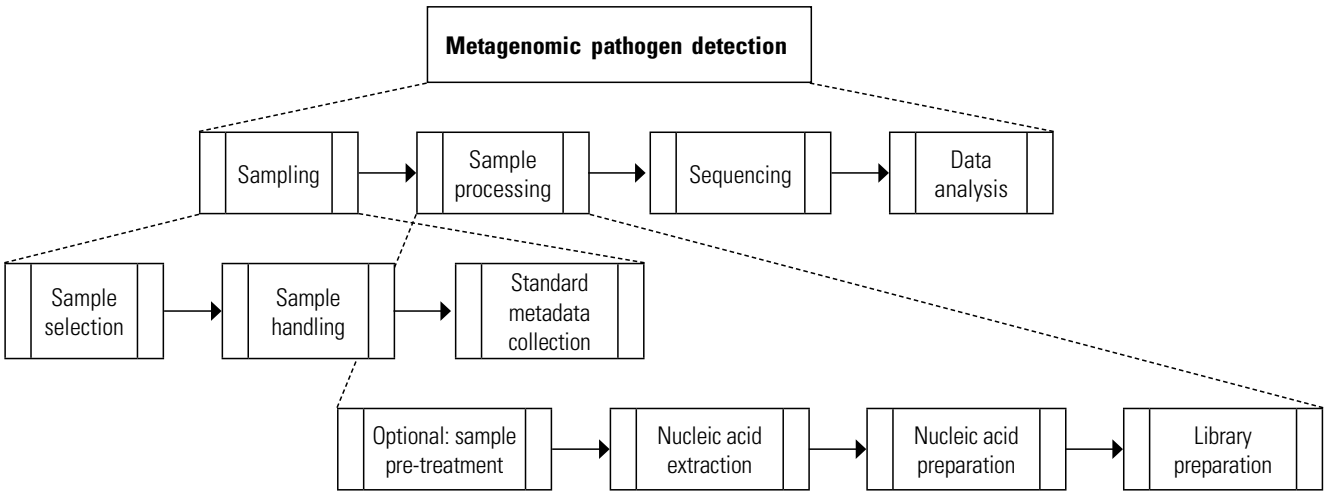


Fig. 1
Overview of the main steps of pathogen identification by metagenomics

Keys to successful pathogen identification

Key determinants for the successful identification of a novel pathogen are:

– the selection of samples
– sample preparation, including nucleic acid extraction
– preparation of the DNA library for sequencing.

Also, with regard to the sequencing platform, several parameters which influence the sensitivity and specificity of the analysis should be taken into account. Finally, data analysis is a major factor in the identification of potential pathogens.

Sampling

The sampling procedure itself has a number of subroutines to consider, namely how to select the animals for sampling, the selection of the actual samples to analyse, sample handling until further processing in the laboratory, and the collection of all relevant metadata.

Sample selection

Sample selection for relevant pathogens may be targeted to animals showing changes in their general condition and displaying clinical signs such as fever, nasal discharge, respiratory signs or even lesions and haemorrhages, in order to enhance the chance of successful analysis. Having identified appropriate target animals, sample selection has two levels. The first level is the choice of the sample material, e.g. blood/serum, saliva, tissue from organs, or excretions. Depending on the pathogen to be identified, the most suitable sample materials can differ. Because different pathogens have different tropisms, they can be more readily detected according to these preferences. For viruses which cause viraemia, the sample of choice might be serum, as may also be the case for other pathogens which circulate in the blood. However, in cases of SBV or other orthobunyaviruses, which only cause very brief viraemia, the time point of sampling is crucial for the success of detection. Table I shows RT-qPCR quantification of viral RNA for different samples from different host species infected with SBV. Obviously, there are huge differences in the viral loads detected, with differences of up to nearly 1,000-fold in different samples from cattle. Owing to the fact that, when first identifying pathogens by metagenomics, nothing is known about the pathogen, it makes sense to analyse a panel of different samples, and also pooled samples, to enhance the likelihood of detection.

The second level of sample selection is the choice of the nucleic acid to target, i.e. DNA or RNA. In this regard, RNA is to be preferred over DNA because replicating infectious agents will produce RNA during gene expression. This will enable the detection of both DNA and RNA viruses. Hence, the chance of identifying a pathogen is highest when targeting RNA. Of course, targeting both RNA and DNA is possible, either separately or in combination, but it involves more effort and higher costs with limited improvement in sensitivity.

Table I
Results for Schmallenberg virus and β-actin reverse transcriptase quantitative polymerase chain reaction from different hosts and sample materials

Host	Sample	Cq (SBV)	Cq (β-actin)
Cattle	Serum	35.57	30.22
Cattle	Serum	28.81	29.52
Cattle	Serum	27.76	29.56
Cattle	Serum	31.33	28.24
Cattle	Serum	36.54	31.09
Cattle	Spleen	31.88	20.56
Cattle	Mandibular LN	33.32	20.81
Cattle	Mesenteric LN	33.08	23.91
Sheep	Serum	28.32	28.35
Sheep	Serum	28.42	31.37
Sheep	Serum	27.42	28.43
Sheep	Serum	37.23	32.34
Sheep	Serum	27.88	31.71
Sheep	Spleen	32.96	24.66
Sheep	Spleen	34.27	26.11
Sheep	Mesenteric LN	35.61	22.55
Sheep	Mesenteric LN	32.45	21.49
Sheep	Mesenteric LN	35.19	22.69
Sheep	Peyer's patches	35.05	21.72
Mouse	Spleen	29.21	26.05
Mouse	Spleen	36.24	26.50
Mouse	Liver	33.18	26.94
Mouse	Liver	36.24	26.50
Mouse	Liver	21.27	25.12
Mouse	Liver	21.43	24.47

Cq: quantification cycle
LN: lymph node
SBV: Schmallenberg virus

Sample handling

Sample handling is another very important issue. Because improper sample handling and storage may bias the results, care has to be taken after sampling. Improper storage, for

example at elevated temperatures, may allow microbial growth. This might bias the results towards aerobic bacteria which will be favoured by the storage conditions. Another factor which may bias the results, or even make the detection of the causative agent impossible, is nucleic acid degradation by nucleases present in the sample, either naturally within the sample or from multiplication of microbes in the sample. Therefore, the samples should be handled with care and processed as quickly as possible after sampling. Finally, sample contamination must be avoided.

Standard metadata collection

To be able to relate clinical signs to the detected species, and for the analysis of potential multi-factorial diseases, it is necessary to collect sufficient metadata. For instance, the clinical course may differ depending on the animal's age or weight, its vaccination status, husbandry or feeding. Vector-borne transmission may occur only within a certain time period (season) in the year. Previous infections may also influence the clinical course. In addition, a comprehensive set of metadata may also help to fulfil Koch's postulates as adapted to pathogen detection by metagenomics (see below).

Sample processing

As for sampling, in sample processing there are also a number of important parameters which have a strong impact on the final results. Among them are the accessibility of the nucleic acids to the sample preparation procedures and the possibility of focusing the sequencing effort by molecular (capture, amplification) or other enrichment techniques.

Sample pre-treatment

Procedures that aim to focus the sequencing on non-host RNA can be included in the sample preparation procedure; however, this may also bias the results. For instance, centrifugation and/or filtration of the sample may result in loss of bacteria and eukaryotic parasites while helping to focus on viral pathogens. Therefore, when attempting to identify an unknown pathogen, care should be taken not to exclude certain groups of pathogens during sample preparation. Several procedures have been proposed for targeting the sequencing effort to viruses. These include SISPA (sequence-independent, single-primer amplification) (12) and SISPA-derived methods (13, 14), VIDISCA methods (virus discovery based on complementary DNA/amplified fragment length polymorphism) (15), and the TUViD-VM method (tissue-based unbiased virus detection for viral metagenomics) (16). With the exception of the TUViD-VM method, all these protocols are optimised for processing liquid samples. All the protocols include centrifugation or filtration procedures which cause loss of bacterial and eukaryotic parasite cells and, therefore, are not suitable for detection of bacterial or parasitic pathogens.

Another issue for some of these enrichment techniques is that the nuclease digestion included in the procedures may lead to information loss caused by degradation of unprotected nucleic acids of the pathogens of interest. Therefore, if the sample has not been handled properly, the necessary information contained in the nucleic acids will be degraded. The detailed data presented by Kohl *et al.* (16) imply that their protocol works specifically with enveloped RNA viruses because the proportion of poxvirus-specific sequences is significantly (Fisher's exact test for count data: $p \leq 2.2e^{-16}$) reduced (six-fold) and the reovirus they used as a representative of non-enveloped viruses was not detected at all. This example clearly demonstrates the problems that may be caused by extensive manipulation of the samples.

Additionally, data published by Rosseel *et al.* (17) on the comparison of different sample pre-treatments for viral metagenomics show that the random PCR included in some protocols generates a substantial amount of presumably artificial sequences for which no match can be detected in the databases, especially when the amount of input nucleic acids is low. In all cases, random PCR caused a statistically significant (Fisher's test: $p \leq 0.0278$) decrease in the proportion of Newcastle disease virus-specific reads. Moreover, their data also show that rRNA depletion can be deleterious when attempting to measure the content of virus-specific RNAs. This was the case for most serum samples in that study, for which the virus/host ratio decreased significantly (Fisher's test: $p \leq 0.0143$) after rRNA depletion. Only a combination of rRNA depletion with a DNase digest of nucleic acids isolated from serum or tissue samples resulted in a significant positive effect (Fisher's test: $p \leq 3.12e^{-13}$). Taken together, the examples presented show that care must be taken with procedures that aim to enrich specific fractions of the sample nucleic acids.

Nucleic acid extraction

An important parameter related to pathogen detection is the accessibility of the pathogen's nucleic acid. Given that nucleic acids may be protected by robust bacterial cell walls or – in the case of viruses, intracellular bacteria and eukaryotic parasites – by being encased in solid host tissues, the technique of nucleic acid extraction must ensure the proper disintegration of host, bacterial and eukaryotic parasite cells as well as viruses in order to access the nucleic acids. Therefore, a method of mechanical disruption which will break all cellular structures is necessary. However, this disruption should not unduly affect the nucleic acids, to ensure their availability for sequencing. After homogenisation of the sample, nucleic acids need to be extracted; various protocols and commercial kits exist for this. An important point to consider is whether to divide the sample into two in order to extract both DNA and RNA as individual samples in parallel. For the RNA sample, a DNase digest is necessary when preparing the DNA library

from the input RNA, to prevent an excess of DNA from masking the RNA sequences.

Nucleic acid preparation

In the case of RNA, after extraction of the nucleic acids, synthesis of cDNA and, because most library preparations are optimised for double-stranded (ds)DNA, second-strand synthesis are necessary. Similarly, second-strand synthesis may also be required if using DNA as the input material. Failure to use dsDNA as the input may result in loss of viruses with single-stranded (ss)DNA genomes, such as species from the families *Circoviridae* and *Anelloviridae*. In order for the overall procedure to achieve sufficient sensitivity, the input amount should not fall below a certain threshold. It must be sufficient to ensure that, among the host nucleic acids, pathogen genome fragments are present. An input amount that is too low cannot be compensated for by random amplification (whole-genome amplification [WGA] or whole-transcriptome amplification [WTA]) of the input material or by PCR amplification of the library, because missing pathogen information will not be generated by any amplification. Rather, redundancy of the information is generated, and therefore the complexity of the library does not reflect the complexity of the sample (3), the whole procedure becomes biased (18, 19, 20, 21, 22) and random sequences may be generated (23, 24). DNA generated artificially during amplification steps can lead to a high proportion of unclassifiable sequences (17).

Library preparation

Library preparation is specifically designed for the respective sequencing platforms. Nonetheless, there are some essential points to consider in all cases. First, the protocol must fit the available input nucleic acids. Some protocols are designed to work with minute amounts of input DNA or RNA; others require several micrograms of input material. The procedure for the library preparation can affect the uniformity of the sequencing depth. For instance, the library preparation procedure using 'tagmentation' (a process in which DNA is fragmented enzymatically and sequencing adapters are mounted simultaneously to the DNA fragments to yield a functional library) was shown to have a guanine/cytosine (GC)-bias (25, 26). Optimisation of library preparation may be necessary with regard to DNA fragmentation and minimisation of purification steps in order to reduce sample loss and prevent cross-contamination.

Sequencing

The choice of a suitable sequencing platform for metagenomic pathogen identification is an important matter. In this regard, throughput, as measured by the number of reads, length of the generated reads, quality of the reads and carry-over between sequencing runs, as well as the runtime itself, have to be taken into account.

The total number of reads is important because the ratio of pathogen and host nucleic acids, i.e. the amount of pathogen nucleic acids within the complete pool of nucleic acids, may be unfavourable for the detection of pathogen sequences if the total size of the dataset is too small. The probability of detecting at least one pathogen read depends on the proportion of pathogen molecules in a given nucleic acid preparation and the sample size taken from this nucleic acid pool. For instance, sequencing 100,000 molecules from a nucleic acid pool of 10,000,000 molecules containing ten pathogen molecules (similar to the situation for swab samples in [11]) will enable the detection of at least one pathogen sequence with a probability of 9.5%. Sequencing 1,000,000 molecules from this library may be sufficient because the probability of detecting at least one pathogen read will rise to 63.2%, and analysing 10,000,000 molecules will result in a probability of 99.995%.

The composition of the library is determined by the mass ratio of pathogen and host nucleic acids, i.e. the total mass of pathogen and host nucleic acids, respectively, in the nucleic acid preparation used for library construction. The two parameters that influence this mass ratio, and hence the probability of pathogen detection, are the copy number and the size of each of the copies. For instance, one copy of the SBV genome corresponds to three molecules with a total of 12.1 kilobases. One copy of the cattle transcriptome comprises 5,400 molecules (27), not considering tissue-specific gene expression, which accounts for up to approximately 7,000 additional transcripts (27). Under the assumption that the median length of bovine messenger RNA (mRNA) is 2,500 bases (as calculated from the bovine mRNA sequences available in the National Center for Biotechnology Information [NCBI] refseq database as of 1 July 2015) and disregarding the high copy number rRNA molecules, one copy of the host transcriptome corresponds to 13.5 megabases (without tissue-specific transcripts). Hence, there is a ratio of one base of pathogen nucleic acids to 1,115 bases of host nucleic acids, which equals one pathogen read in 1,115 host reads.

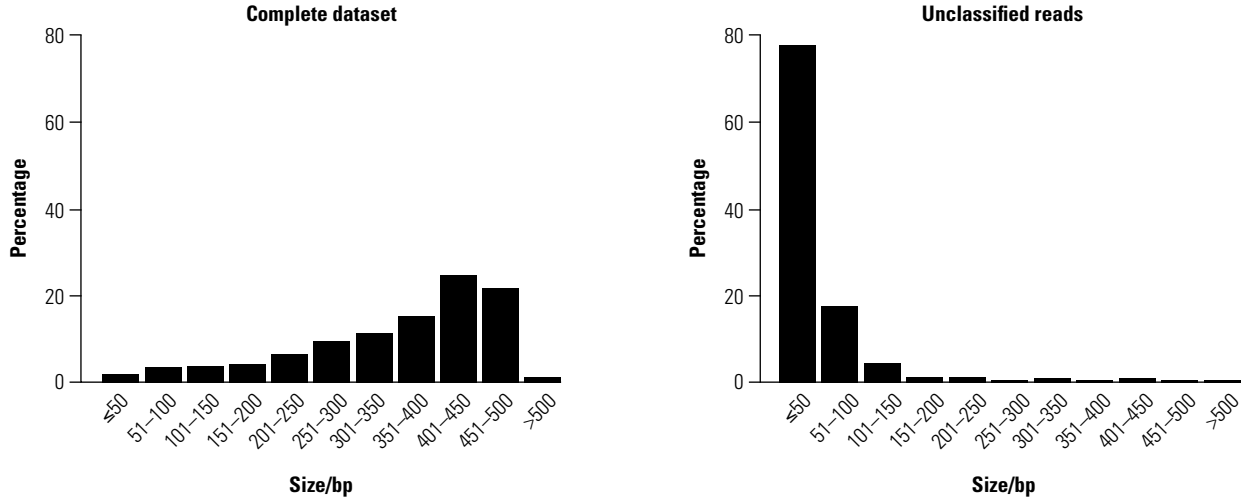
The situation is in fact even more challenging, because the true ratios of pathogen and host nucleic acids must also be considered. Hence, the ratio of one copy of the SBV genome to one copy of the host transcriptome, as assumed above, is not valid: lower copy numbers for SBV in serum (up to 32-fold, or 5 quantification cycles [Cq], see Table I) must be applied. This finally results in a representation of SBV of about one read within 3.6×10^4 reads. In a tissue sample with a median Cq difference of 11 (Table I), i.e. a 2,048-fold under-representation of SBV, only one SBV-specific read in a total of 2.3×10^6 reads may be expected, again disregarding rRNA. Sequencing a sample from a library with these characteristics would find at least one pathogen read with a probability of not more than 0.4% (sample size 10,000 reads), 4.3% (sample size 100,000 reads) or

35.3% (sample size 1,000,000 reads). Although the relative abundances are only very rough estimates, they highlight the problem of masking of the pathogen by host sequences if the sample size is too small. The orders of magnitude of the differences in the abundances are confirmed by data presented in various publications (11, 16, 17, 28) and apply even after host sequence depletion (17, 28).

The length of the generated reads is another determinant of metagenomic pathogen detection because the power of the algorithms used for read classification depends directly on read length. If the reads are too short, it may not be possible to generate a significant and specific alignment. Figure 2 shows the read length distribution of the initial SBV dataset (4) and the reads remaining unclassified after analysis with the RIEMS (Reliable Information Extraction from Metagenomic Sequence datasets) software (29). This clearly illustrates that the use of shorter reads is more problematic and that longer reads may be more useful for pathogen detection and also identification, because the specificity of the classification of short sequences is lower. Comparison of a classification of the reads from the initial dataset that led to the identification of SBV (4) with the same reads shortened to approximately 100 base pairs clearly demonstrates that read length is important for pathogen identification using metagenomics. The original mean read length was 315 bases, and the artificially shortened reads had a mean length of 96 bases. Owing to this shortening, the total number of reads that could not be classified nearly doubled, from 4.7% to 7.7%. More importantly, the number of reads that could be assigned to the orthobunyaviruses decreased from seven to one, making identification nearly impossible. Another analysis of the impact of read length on the error rate of sequence classification was reported by Bibby *et al*. (30), who found that, in particular, the rate of misclassifications by BLASTx and tBLASTx (Basic Local Alignment Search Tools) increased dramatically with shorter reads. Moreover, for a robust classification, the error rate of the sequencer should be as low as possible because otherwise, depending on the software, the similarity values may be too low to detect similar sequences (29) or reads may be misclassified, thereby masking the genetic information of the pathogen.

Ensuring the correct sorting of reads into the respective datasets may become a problem when using equipment multiple times for library preparation, sequencing run setup, or the sequencing itself. For instance, the Illumina® MiSeq instrument always receives samples via the same port with fixed tubing. This causes a substantial carry-over between runs (31); according to the manufacturer's documentation, carry-over in a typical run may be up to 0.1% (32). This problem can be reduced to a certain extent by ensuring that the same molecular barcode is not used in subsequent runs. However, this will not be possible in all settings and will, in addition, not completely resolve the problem because of index misassignment (where a cluster is assigned an incorrect index sequence) of up to 0.06% (31). This needs to be considered when dealing with datasets with very low abundance of potential pathogen reads.

Large differences exist between different platforms in the runtime of the sequencing. Given that rapid results may be important in diagnostic settings, the runtime should be taken into account during platform selection. While the Illumina® and the SOLiD® instruments have runtimes in the range of days before the final sequence reads are available



bp: base pairs

Fig. 2
Read length distributions of the complete dataset in which Schmallenberg virus sequences were initially identified (left) and from the reads that could not be classified (right)

(33), sequencing with the Ion Torrent™ PGM platform only requires hours from starting the run until availability of the sequencing reads. However, additional time for further sample preparation steps (emulsion PCR of the libraries and subsequent enrichment of template-positive Ion spheres) and chip loading have to be considered.

Data analysis in metagenomics

In general, for metagenomic identification of pathogens, analysis of the sequencing data is aimed initially at the taxonomic classification of the reads obtained. Identification of similar sequences in the nucleotide databases, for instance by BLAST (34), Kraken (35), FASTA (36) or other available algorithms or workflows, is the first step. This can be challenging because the generated datasets may comprise millions or even billions of reads that need to be classified. This is computationally demanding, and the results of the similarity searches need to be screened for the relevant information. For BLAST results this is frequently done using MEGAN (37), which summarises the results of BLAST analyses. In order to obtain manageable datasets that limit the subsequent filtering for relevant information, and to reduce the necessary computing, other approaches limit the reference databases (38, 39). In this case, however, the detected pathogens are defined by the references provided by the user and, therefore, the analysis will be biased. Moreover, there is no possibility of detecting unexpected pathogens. In addition, most available tools are centred upon the analysis of datasets of human origin. This may decrease the efficiency of the analysis of veterinary datasets.

Given that unbiased screening of the full dataset against the complete database of known sequences requires a high computing capacity, in some instances researchers only screen the sequence datasets against partial databases, for instance those containing only viral sequences. In some cases this may be combined with tBLASTx, the BLAST algorithm set to search a translated nucleotide sequence query in a translated nucleotide sequence database, i.e. aligning two nucleotide sequences according to the alignment of the deduced amino acid sequences. This reduces the stringency of the search and results in a higher number of hits. However, the reliability of a tBLASTx-search in a viral database is significantly lower than that of screening a complete nucleotide database by BLASTn (30). (BLASTn is the BLAST algorithm used to search for a nucleotide sequence within a nucleotide sequence database; it is slower than MegaBLAST but it is more sensitive.) This can also be shown by analysis of random sequences. The tBLASTx analysis of 1,000 random sequences, each comprising 250 bases, when used to search the complete NCBI nucleotide sequence database, yielded two hits, although not with pathogens, with expectation values (e-values) ≤ 0.0002 and identities for query and subject sequences of ≤ 52%. The same analysis using only a virus nucleotide sequence database, as frequently performed (40, 41, 42), yielded a hit with a porcine bocavirus with an e-value of 0.002 and a similarity value for the query and subject sequences of 41%. In contrast, when searching against the complete NCBI nucleotide sequence database, neither MegaBLAST nor BLASTn searches of the same random sequences yielded significant hits.

In summary, state-of-the-art data analysis requires an unbiased approach to screening for the full taxonomic content of the sample, in order to meet the requirements of the unbiased sequence datasets generated by NGS. Unbiased analysis of complete datasets with the classification of all individual reads is, for instance, possible with the recently published RIEMS workflow (29).

Follow-up after pathogen identification

After identification of a novel pathogen the association of the agent and the clinical symptoms needs to be confirmed. The first criteria to unequivocally link a pathogen with clinical symptoms were proposed by Koch in 1890 (43). The criteria set forth in that presentation were subsequently discussed and adapted (44, 45, 46). It was also proposed that identification of sequences related to the same pathogen from individuals with equivalent symptoms be sufficient as a first proof (44, 45); moreover, a scoring system to assess the certainty of the causality has been proposed (47). Nevertheless, Koch's postulates remain the gold standard for establishing a causal link between pathogen and disease.

In order to prove the association of the potential pathogen with the observed clinical signs according to Koch's postulates, it is necessary to have a pure culture or an isolate. Information gained by sequencing can assist in this. Subsequently, for in-depth characterisation, nucleic acids from isolates can be used to generate full-genome sequences for viruses, or draft genomes comprising the full set of pathogenicity determinants may be generated for bacterial or eukaryotic pathogens. The initial sequence information will also be the basis for designing the first diagnostic real-time PCR assays to enable the screening of further samples to assess the spread of the pathogen in an animal or within a population. This may also help to fulfil the requirement of showing a strong association of the presumed pathogen with the disease, according to the aforementioned criteria.

Conclusions

Modern sequencing technologies allow, for the first time, a metagenomic approach for pathogen identification

during outbreaks and pathogen discovery by targeted or untargeted screening. NGS-based virus discovery is a good example to use to demonstrate the impressive power of the new technologies, but also the pitfalls.

A reasonable metagenomic approach based on NGS must consider the best sample materials, optimised sample handling and processing procedures, and suitable data analysis pipelines. These pipelines should allow screening of the full dataset against a complete sequence database in order to avoid a bias in read allocation.

Acknowledgements

The authors thank Kerstin Wernike for providing the SBV real-time PCR data and Ariane Belka for help in the calculation of probabilities for pathogen detection. The study has been funded in part by the European Union Horizon 2020 program (European Commission Grant Agreement No. 643476 'COMPARE').

Les méthodes d'identification des agents pathogènes basées sur la métagénomique

D. Höper, T.C. Mettenleiter & M. Beer

Résumé

Depuis l'avènement des technologies de séquençage de nouvelle génération, le criblage non ciblé d'échantillons prélevés au cours d'un foyer de maladie afin d'identifier l'agent pathogène en recourant à la métagénomique est devenu accessible aux plans technique et économique. Néanmoins, un certain nombre d'aspects restent à élucider afin d'exploiter pleinement les possibilités offertes par le séquençage de nouvelle génération pour déceler des virus précédemment inconnus. Les auteurs résument ces aspects pour chaque étape déterminante, depuis le choix des échantillons jusqu'à leur manipulation et traitement, et du séquençage à l'analyse des données, en mettant l'accent sur les difficultés éventuelles.

Mots-clés

Découverte de virus – Logiciel d'analyse – Manipulation d'échantillons – Métagénome – Séquençage de nouvelle génération – Traitement d'échantillons.

Métodos de metagenómica para identificar agentes infecciosos

D. Höper, T.C. Mettenleiter & M. Beer

Resumen

Desde el advenimiento de las técnicas de secuenciación de próxima generación, el cribado no selectivo de muestras tomadas durante un brote para identificar al patógeno empleando la metagenómica ha pasado a ser técnica y económicamente viable. Sin embargo, hay una serie de aspectos que conviene tener en cuenta a fin de poder aprovechar plenamente las posibilidades que ofrecen esas técnicas para descubrir virus. Los autores resumen esos aspectos en relación con las principales etapas que tienen una influencia importante, desde la selección

hasta la manipulación y el procesamiento de las muestras, pasando por la secuenciación y el análisis de datos, haciendo especial hincapié en sus posibles inconvenientes.

Palabras clave
Descubrimiento de virus – Manipulación de muestras – Metagenoma – Procesamiento de muestras – Secuenciación de próxima generación – *Software* de análisis.

References

1. Olsen G.J., Lane D.J., Giovannoni S.J., Pace N.R. & Stahl D.A. (1986). – Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.*, **40**, 337–365. doi:10.1146/annurev.mi.40.100186.002005.

2. Handelsman J., Rondon M.R., Brady S.F., Clardy J. & Goodman R.M. (1998). – Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, **5** (10), R245–R249.

3. Head S.R., Komori H.K., LaMere S.A., Whisenant T., Van Nieuwerburgh F., Salomon D.R. & Ordoukhanian P. (2014). – Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, **56** (2), 61–64, 66, 68, passim. doi:10.2144/000114133.

4. Hoffmann B., Scheuch M., Höper D., Jungblut R., Holsteg M., Schirrmeier H., Eschbaumer M., Goller K.V., Wernike K., Fischer M., Breithaupt A., Mettenleiter T.C. & Beer M. (2012). – Novel orthobunyavirus in cattle, Europe, 2011. *Emerg. Infect. Dis.*, **18** (3), 469–472. doi:10.3201/eid1803.111905.

5. Beer M., Conraths F.J. & van der Poel W.H. (2013). – 'Schmallenberg virus': a novel orthobunyavirus emerging in Europe. *Epidemiol. Infect.*, **141** (1), 1–8. doi:10.1017/S0950268812002245.

6. Bouzalas I.G., Wüthrich D., Walland J., Drögemüller C., Zurbriggen A., Vandevelde M., Oevermann A., Bruggmann R. & Seuberlich T. (2014). – Neurotropic astrovirus in cattle with nonsuppurative encephalitis in Europe. *J. Clin. Microbiol.*, **52** (9), 3318–3324. doi:10.1128/JCM.01195-14.

7. Hause B.M., Ducatez M., Collin E.A., Ran Z., Liu R., Sheng Z., Armien A., Kaplan B., Chakravarty S., Hoppe A.D., Webby R.J., Simonson R.R. & Li F. (2013). – Isolation of a novel swine influenza virus from Oklahoma in 2011 which is distantly related to human influenza C viruses. *PLoS Pathog.*, **9** (2), e1003176. doi:10.1371/journal.ppat.1003176.

8. Ducatez M.F., Pelletier C. & Meyer G. (2015). – Influenza D virus in cattle, France, 2011–2014. *Emerg. Infect. Dis.*, **21** (2), 368–371. doi:10.3201/eid2102.141449.

9. Tong S., Li Y., Rivailler P., Conrardy C., Castillo D.A., Chen L.M., Recuenco S., Ellison J.A., Davis C.T., York I.A., Turmelle A.S., Moran D., Rogers S., Shi M., Tao Y., Weil M.R., Tang K., Rowe L.A., Sammons S., Xu X., Frace M., Lindblade K.A., Cox N.J., Anderson L.J., Rupprecht C.E. & Donis R.O. (2012). – A distinct lineage of influenza A virus from bats. *Proc. Natl Acad. Sci. USA*, **109** (11), 4269–4274. doi:10.1073/pnas.1116200109.

10. Tong S., Zhu X., Li Y., Shi M., Zhang J., Bourgeois M., Yang H., Chen X., Recuenco S., Gomez J., Chen L.M., Johnson A., Tao Y., Dreyfus C., Yu W., McBride R., Carney P.J., Gilbert A.T., Chang J., Guo Z., Davis C.T., Paulson J.C., Stevens J., Rupprecht C.E., Holmes E.C., Wilson I.A. & Donis R.O. (2013). – New World bats harbor diverse influenza A viruses. *PLoS Pathog.*, **9**, e1003657. doi:10.1371/journal.ppat.1003657.

11. Hoffmann B., Tappe D., Höper D., Herden C., Boldt A., Mawrin C., Niederstraßer O., Müller T., Jenckel M., van der Grinten E., Lutter C., Abendroth B., Teifke J., Cadar D., Schmidt-Chanasit J., Ulrich R.G. & Beer M. (2015). – A variegated squirrel bornavirus associated with fatal human encephalitis. *N. Engl. J. Med.*, **373**, 154–162. doi:10.1056/NEJMoa1415627.

12. Reyes G.R. & Kim J.P. (1991). – Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molec. Cell. Probes*, **5** (6), 473–481. doi:10.1016/S0890-8508(05)80020-9.

13. Froussard P. (1993). – rPCR: a powerful tool for random amplification of whole RNA sequences. *PCR Meth. Appl.*, **2**, 185–190. doi:10.1101/gr.2.3.185.

14. Froussard P. (1992). – A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res.*, **20** (11), 2900. doi:10.1093/nar/20.11.2900.

15. De Vries M., Deijs M., Canuti M., van Schaik B.D., Faria N.R., van de Garde M.D., Jachimowski L.C., Jebbink M.F., Jakobs M., Luyf A.C., Coenjaerts F.E., Claas E.C., Molenkamp R., Koekkoek S.M., Lammens C., Leus F., Goossens H., Ieven M., Baas F. & van der Hoek L. (2011). – A sensitive assay for virus discovery in respiratory clinical samples. *PLoS One*, **6** (1), e16118. doi:10.1371/journal.pone.0016118.

16. Kohl C., Brinkmann A., Dabrowski P.W., Radonić A., Nitsche A. & Kurth A. (2015). – Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.*, **21** (1), 48–57. doi:10.3201/eid2101.140766.

17. Rosseel T., Ozhelvaci O., Freimanis G. & Van Borm S. (2015). – Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *J. Virol. Meth.*, **222**, 72–80. doi:10.1016/j.jviromet.2015.05.010.

18. Pinard R., de Winter A., Sarkis G.J., Gerstein M.B., Tartaro K.R., Plant R.N., Egholm M., Rothberg J.M. & Leamon J.H. (2006). – Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216. doi:10.1186/1471-2164-7-216.

19. Aird D., Ross M.G., Chen W.S., Danielsson M., Fennell T., Russ C., Jaffe D.B., Nusbaum C. & Gnirke A. (2011). – Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12** (2), R18. doi:10.1186/gb-2011-12-2-r18.

20. Abbai N.S., Govender A., Shaik R. & Pillay B. (2012). – Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. *Molec. Biotechnol.*, **50** (1), 39–48. doi:10.1007/s12033-011-9412-8.

21. Kim K.H. & Bae J.W. (2011). – Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.*, **77** (21), 7663–7668. doi:10.1128/AEM.00289-11.

22. Sujayanont P., Chininmanu K., Tassaneetrithep B., Tangthawornchaikul N., Malasit P. & Suriyaphol P. (2014). – Comparison of phi29-based whole genome amplification and whole transcriptome amplification in dengue virus. *J. Virol. Meth.*, **195**, 141–147. doi:10.1016/j.jviromet.2013.10.005.

23. Hutchison 3rd C.A., Smith H.O., Pfannkoch C. & Venter J.C. (2005). – Cell-free cloning using phi29 DNA polymerase. *Proc. Natl Acad. Sci. USA*, **102** (48), 17332–17336. doi:10.1073/pnas.0508809102.

24. Qiagen (2011). – REPLI-g® Mini/Midi Handbook. Qiagen, Hilden, Germany.

25. Marine R., Polson S.W., Ravel J., Hatfull G., Russell D., Sullivan M., Syed F., Dumas M. & Wommack K.E. (2011). – Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl. Environ. Microbiol.*, **77** (22), 8071–8079. doi:10.1128/AEM.05610-11.

26. Parkinson N.J., Maslau S., Ferneyhough B., Zhang G., Gregory L., Buck D., Ragoussis J., Ponting C.P. & Fischer M.D. (2012). – Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res.*, **22** (1), 125–133. doi:10.1101/gr.124016.111.

27. Harhay G.P., Smith T.P., Alexander L.J., Haudenschild C.D., Keele J.W., Matukumalli L.K., Schroeder S.G., Van Tassell C.P., Gresham C.R., Bridges S.M., Burgess S.C. & Sonstegard T.S. (2010). – An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome Biol.*, **11** (10), R102. doi:10.1186/gb-2010-11-10-r102.

28. Rosseel T., Scheuch M., Höper D., De Regge N., Caij A.B., Vandenbussche F. & Van Borm S. (2012). – DNase SISPA-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe. *PLoS One*, **7** (7), e41967. doi:10.1371/journal.pone.0041967.

29. Scheuch M., Höper D. & Beer M. (2015). – RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC Bioinformatics*, **16**, 69. doi:10.1186/s12859-015-0503-6.

30. Bibby K., Viau E. & Peccia J. (2011). – Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Lett. Appl. Microbiol.*, **52** (4), 386–392. doi:10.1111/j.1472-765X.2011.03014.x.

31. Nelson M.C., Morrison H.G., Benjamino J., Grim S.L. & Graf J. (2014). – Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One*, **9** (4), e94249. doi:10.1371/journal.pone.0094249.

32. Illumina (2013). – Reducing run-to-run carryover on the MiSeq using dilute sodium hypochlorite solution. Illumina, San Diego, California.

33. Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P. & Gu Y. (2012). – A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341. doi:10.1186/1471-2164-13-341.

34. Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990). – Basic Local Alignment Search Tool. *J. Molec. Biol.*, **215** (3), 403–410. doi:10.1016/S0022-2836(05)80360-2.

35. Wood D.E. & Salzberg S.L. (2014). – Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15** (3), R46. doi:10.1186/gb-2014-15-3-r46.

36. Pearson W.R. & Lipman D.J. (1988). – Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85** (8), 2444–2448.

37. Huson D.H., Auch A.F., Qi J. & Schuster S.C. (2007). – MEGAN analysis of metagenomic data. *Genome Res.*, **17** (3), 377–386. doi:10.1101/gr.5969107.

38. Naeem R., Rashid M. & Pain A. (2013). – READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics*, **29** (3), 391–392. doi:10.1093/bioinformatics/bts684.

39. Bhaduri A., Qu K., Lee C.S., Ungewickell A. & Khavari P.A. (2012). – Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*, **28** (8), 1174–1175. doi:10.1093/bioinformatics/bts100.

40. Sachsenröder J.T.S., Hammerl J.A., Janczyk P., Wrede P., Hertwig S. & Johne R. (2012). – Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing. *PLoS One*, **7** (4), e34631. doi:10.1371/journal.pone.0034631.

41. Reyes A., Haynes M., Hanson N., Angly F.E., Heath A.C., Rohwer F. & Gordon J.I. (2010). – Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, **466** (7304), 334–338. doi:10.1038/nature09199.

42. Pérez-Brocal V., García-López R., Vázquez-Castellanos J.F., Nos P., Beltrán B., Latorre A. & Moya A. (2013). – Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin. Transl. Gastroenterol.*, **4**, e36. doi:10.1038/ctg.2013.9.

43. Koch R. (1890). – Ueber bakteriologische Forschung [About bacteriological research]. *In* Proc. 10th International Medical Congress, Berlin, 1890. Verlag von August Hirschwald, Berlin, 35–47.

44. Fredricks D.N. & Relman D.A. (1996). – Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.*, **9** (1), 18–33.

45. Mokili J.L., Rohwer F. & Dutilh B.E. (2012). – Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.*, **2** (1), 63–77. doi:10.1016/j.coviro.2011.12.004.

46. Rivers T.M. (1937). – Viruses and Koch's postulates. *J. Bacteriol.*, **33** (1), 1–12.

47. Lipkin W.I. (2013). – The changing face of pathogen discovery and surveillance. *Nat. Rev. Microbiol.*, **11** (2), 133–141. doi:10.1038/nrmicro2949.