METHOD DEVELOPMENT

Discovery and diagnosis of new viral pathogens: proposal for a generic workflow based on next-generation sequencing and new integrated data analysis approaches

Dirk Höper*, Anne Pohlmann**.

Abstract

Diagnostic metagenomic analyses gained more impact for pathogen detection and discovery in recent years due to increasing opportunities for next-generation sequencing at simultaneously decreasing prices. However, as with all novel technologies, there is a lack of standardisation in this field. But, for day-to-day routine use in a diagnostic laboratory, standardisation is urgently required. This standardisation has to take into account all steps from sampling through library preparation to data evaluation. In this short outline, some examples of metagenomics-based pathogen identification are highlighted and the most critical steps of the whole procedure are discussed. Lastly, the reader is pointed to currently ongoing initiatives that aim to achieve these goals of standardisation to pave the way for broader use of next-generation sequencing in diagnosis.

Keywords

- Data analysis
- Diagnosis
- Metagenomics
- Next-generation sequencing
- Virus

** FLI, Institute of Diagnostic Virology, 17493, Greifswald-Insel Riems, Germany

* Corresponding author : dirk.hoeper@fli.bund.de

METHOD DEVELOPMENT

Introduction

Fast and reliable detection is the basis for the discovery and control of viral infections. The established molecular methods, however, are in most cases only useful for the detection of known pathogens. During recent years, it has become more and more obvious that human and animal health is threatened by new emerging viral infectious diseases. Such diseases may spread very fast due to increasing travel and global trade chains (*e.g.* food) and transport activities. The relevance of novel diseases can easily be illustrated by several new emerging viral pathogens that were discovered over the last few years. The novel orthobunyavirus Schmallenberg virus (SBV) [Hoffmann *et al.*, 2012] or the new MERS coronavirus described in 2012 [Zaki *et al.*, 2012] have already been detected in countries around the world. New variants of influenza viruses regularly cause outbreaks globally, as for example the recent Influenza subtype H7N9 in China in 2013 [Gao *et al.*, 2013], and subtype H5N8 in Europe in 2014 and 2015 [Harder *et al.*, 2015; Verhagen *et al.*, 2015]. Influenza is a perfect example of a zoonotic infectious agent that not only causes heavy losses in poultry and swine production but also endangers human health. Examples of the detection of new zoonotic pathogens, such as the new variegated squirrel 1 bornavirus (VSBV-1) recently detected in variegated squirrels (*Sciurus variegatoides*) [Hoffmann *et al.*, 2015], underline the importance of surveillance and the search for new emerging viral pathogens in animal and human samples, as well as the need for open-view methods for fast and reliable detection and identification.

Workflow approaches

One key approach for the detection of new emerging pathogens is sequencing with next-generation sequencing (NGS) techniques. These techniques enable unbiased sequencing of all nucleic acids in a sample. Despite the fact that technical improvements and a decrease in sequencing costs over recent years have paved the way for wider use, NGS is still rarely used in day-to-day diagnosis. One major barrier is often the lack of comprehensive and easy-to-use harmonised workflows for all steps from sample to final result. Figure 1 depicts the main steps that must be included in such a stratified workflow for successful pathogen detection. This workflow must comprise methods for all steps from sampling, sample processing and sequencing, to bioinformatics tools and finely tuned methods for data analysis.

In addition, all data and the analytical results should be available in an easily comprehensible and user-friendly format. Currently, bioinformatics can still be a barrier for metagenomic pathogen detection. Bioinformatics must address use of huge data sets with respect to long-term storage and availability. More importantly, unbiased analysis which must ensure the identification of the needle in the haystack, *i.e.* finding single sequencing reads of viral origin in the overload of host sequences, is necessary. Importantly, pathogen identification does not end with the detection of sequences since this is only the starting point. For confirmation, conventional virological testing is needed, including all efforts to isolate the new virus as a prerequisite for the fulfilment of Koch's postulates. This is also the basis for further biological characterisation and development of serological diagnostics or prototype vaccines.

Over the last decade, various NGS techniques have been described [Margulies *et al.*, 2005; Bentley *et al.*, 2008; Rothberg *et al.*, 2011] and subsequently developed to commercially available products. These so-called 2nd generation NGS methods enable sequencing of nucleic acids from diverse samples even with limited DNA/RNA availability, as is for instance regularly the case with diagnostic samples. These methods require amplification of the sequencing-ready sample to ensure signal detection during sequencing. The currently available 2nd generation NGS platforms enable the sequencing of millions up to billions of individual DNA molecules simultaneously. Novel NGS of the 3rd generation [Clarke *et al.*, 2009; Eid *et al.*,

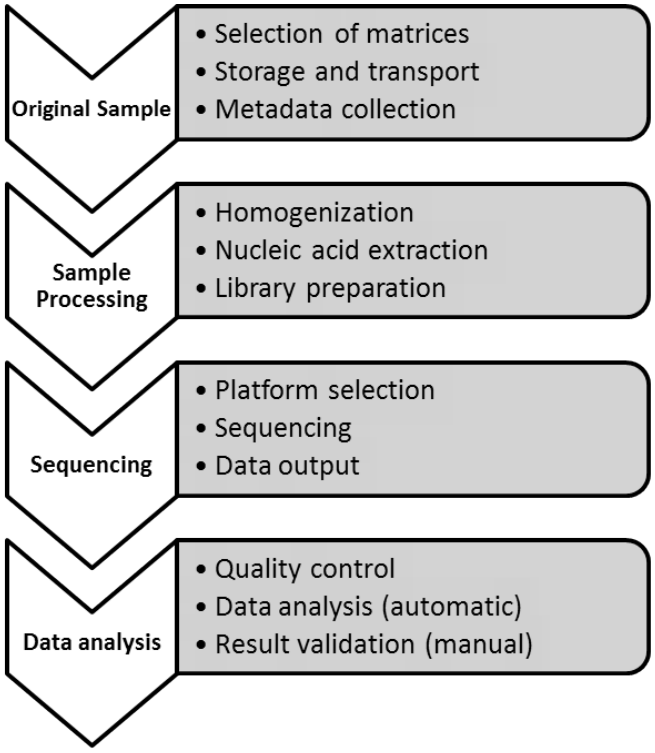
METHOD DEVELOPMENT

2009] requires no sample amplification for detectable signal intensity but suffers from the need for high amounts of input DNA and higher error rates, and might therefore only be instrumental for direct sequencing of samples of sufficient DNA purity and quantity.

Several examples have already proven the successful application of NGS techniques for diagnostic purposes, including discovery of new pathogens [Cox-Foster *et al.*, 2007; Palacios *et al.*, 2008; Hoffmann *et al.*, 2012; Hoffmann *et al.*, 2015], but a closer look at studies reveals that the methods used are far from being comparable, even though they are based on the same principles. This highlights the need to evaluate the different proposed procedures in order to develop a validated harmonised workflow for pathogen detection. Specifically, key steps of the workflow from sampling to sample preparation prior to NGS itself are not always carried out in an optimal way. In addition, not all steps of the procedures found in the literature are well balanced along the complete workflow, hampering the interpretation of NGS results. Key steps in this regard are sampling, including selection of sample matrices, and sample transport, storage and processing. Regarding sample selection, it is obvious that due to their different tropisms and transmission routes, pathogens are naturally not uniformly present in all sample materials like organs, serum, or swabs. Thus the selection of the sample matrix has an enormous impact on the results [Hoffmann *et al.* 2015]. It is therefore of utmost importance to include different matrices in NGS-based diagnostics. In addition, reliable and unbiased data analysis, *i.e.* screening of all sequencing reads against the complete set of available reference sequences and not only against a pre-defined subset, is crucial for any successful diagnostic implementation of NGS for pathogen detection. Most analysis strategies are essentially based on similarity comparison of the sequences gained by NGS with already available information in public databases [Bhaduri *et al.*, 2012; Naeem *et al.*, 2013; Byrd *et al.*, 2014; Scheuch *et al.*, 2015]. These similarity-based approaches limit the sensitivity of the overall process.

Moreover, the quality of the databases used is a decisive criterion for both the sensitivity and the speed of analysis. All data analysis programs/workflows have in common that they are not yet available in a harmonised and easy-to-use environment that integrates sequence data and all other available metadata, like case information with the sequence analysis results. For reliable use in day-to-day diagnosis, it is also essential that the analysis programs are continuously updated and maintained, and are regularly tested and audited.

FIGURE 1 / Principal steps of a workflow for the detection of novel pathogens based on next-generation sequencing.



Scientists and diagnosticians worldwide have recognised that the optimisation of all steps of the workflow and their integrated linkage is the prerequisite for successful implementation of virus discovery and detection by modern NGS techniques in day-to-day clinical diagnosis. Currently, globally coordinated activities dedicated to optimisation and harmonisation of the above-mentioned procedures for NGS-driven pathogen detection and in-depth characterisation are underway. Examples include the “Global Microbial Identifier” initiative (http://www.globalmicrobialidentifier.org/) and the EU-funded project “COMPARE” (http://www.compare-europe.eu/). Both tackle the present shortcomings within

powerful consortia with partly overlapping activities but with slightly different aims and visions. Together, these and other initiatives will make a substantial contribution paving the way for routine use of NGS for next-generation diagnostics.

References

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, ST Ruediger, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.

Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. 2012. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* 28:1174-1175.

Byrd AL, Perez-Rogers JF, Manimaran S, Castro-Nallar E, Toma I, McCaffrey T, Siegel M, Benson G, Crandall KA, Johnson WE. 2014. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC *Bioinformatics* 15:262.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* 4:265-270.

Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui L, Hutchison SK, Simons JF, Egholm M, Pettis JS, Lipkin WI. 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318:283-287.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133-138.

Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W, Chen J, Jie Z, Qiu H, Xu K, Xu X, Lu H, Zhu W, Gao Z, Xiang N, Shen Y, He Z, Gu Y, Zhang Z, Yang Y, Zhao X, Zhou L, Li X, Zou S, Zhang Y, Li X, Yang L, Guo J, Dong J, Li Q, Dong L, Zhu Y, Bai T, Wang S, Hao P, Yang W, Zhang Y, Han J, Yu H, Li D, Gao GF, Wu G, Wang Y, Yuan Z, Shu Y. 2013. Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *New England Journal of Medicine* 368:1888-1897.

Harder T, Maurer-Stroh S, Pohlmann A, Starick E, Höreth-Böntgen D, Albrecht K, Pannwitz G, Teifke J, Gunalan V, Lee RT, Sauter-Louis C, Homeier T, Staubach C, Wolf C, Strebelow G, Höper D, Grund C, Conraths FJ, Mettenleiter TC, Beer M. 2015. Influenza A(H5N8) Virus Similar to Strain in Korea Causing Highly Pathogenic Avian Influenza in Germany. *Emerging Infectious Diseases* 21:860-863.

Hoffmann B, Scheuch M, Höper D, Jungblut R, Holsteg M, Schirrmeier H, Eschbaumer M, Goller KV, Wernike K, Fischer M, Breithaupt A, Mettenleiter TC, Beer M. 2012. Novel Orthobunyavirus in

METHOD DEVELOPMENT

Cattle, Europe, 2011. *Emerging Infectious Diseases* 18:469-472.

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 RG, Beer M. 2015. A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. *New England Journal of Medicine* 373:154-162.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Naeem R, Rashid M, Pain A. 2013. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* 29:391-392.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI. 2008. A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. *New England Journal of Medicine* 358:991-998.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348-352.

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(1):69.

Verhagen JH, van der Jeugd HP, Nolet BA, Slaterus R, Kharitonov SP, de Vries PP, Vuong O, Majoor F, Kuiken T, Fouchier RA. 2015. Wild bird surveillance around outbreaks of highly pathogenic avian influenza A(H5N8) virus in the Netherlands, 2014, within the context of global flyways. *Eurosurveillance* 20(12).

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *New England Journal of Medicine* 367:1814-1820.