Revised version 2

Running Title: Precision pandemic preparedness with metagenomics

Keywords: infection, diagnosis, disease outbreaks, metagenomics.

Title: Precision pandemic preparedness-Improving diagnostics with metagenomics

Authors: Kumeren N. Govender

Affiliations:

Nuffield Department of Medicine, University of Oxford, Oxford, UK (K N Govender)

Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa (K N Govender)

Department of Environmental Health, Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa (K N Govender)

Abstract - (69 words)

The threat posed by novel pandemics in the future remain active. Equipping our routine laboratory with clinical metagenomics to detect unknown threats early on offers a considerable advantage, and may be feasible and scalable with the ability to identify complicated infectious diseases in routine care. Though several technical and regulatory challenges still exist, clinical metagenomics may improve individual patient outcomes, and provide earlier warning signs to improve pandemic preparedness.

Text – (2267 words)

Introduction

Almost all recent pandemics have a viral etiology from animal origin, which pose a great threat with its intrinsic ability for cross-species transmission (1). It is estimated that 1.7 million unknown viral species from viral families exists in mammals and birds, which are the most important reservoir for viral zoonoses, of which 37-49% pose a significant outbreak threat (2). However, if viruses are our enemy, we have limited ways of knowing what the enemy appears to be, at least not before an outbreak of great proportion is already before us. Clinical metagenomic may offer a solution to this not only by identifying complex infectious diseases that are conventionally untestable, but also by rapid identification of novel pathogens, some of which may trigger an urgent outbreak response.

Most notably, the recent identification of SARS-CoV-2 as the causative agent of unknown cluster pneumonia cases seen in Wuhan City, can be seen as a successful use case of metagenomic sequencing for outbreak pathogen detection. Although the earliest known date of symptom onset from the first patient identified was Dec 1, 2019, the causative agent SARS-CoV-2 was identified 37 days later on Jan 7, 2020 by the Chinese Center for Disease Control and Prevention (3). Though much of this delay was recognizing that an outbreak was in fact taking place, the actual metagenomic analysis from bronchoalveolar lavage fluid that determined the identification of a novel coronavirus was performed rapidly in a matter of hours to days (4). Although rapid and widespread metagenomic testing may have identified this pathogen earlier, there are several hurdles to large-scale implementation. Metagenomic approaches are still largely heterogenous where there is little convergence on best practices, which make automation of this technology particularly challenging, not to mention the time-consuming processes and high costs associated with its operation. However, there has been significant scientific progress in the field, and with the development of reporting specifications including the STROBE-metagenomics

guidelines, metagenomics is positioned to be more widely used and accepted (5). Once effective pipelines from sample to report are automated, unusual and routinely untestable infections may be detected within hours or days of initial presentation, often benefitting selected patients and potentially offering early warning signs for pathogens with pandemic potential.

Detecting the invisible enemy

Clinical metagenomics is the application of sequencing all genes in a single clinical sample without the need for isolation or lab cultivation. These sequences are then bioinformatically classified using a comprehensive database of known microorganism DNA, producing phylogenetic matches to the most closely related genus and species. Had this technology been widely accessible and available at the bedside, the novel SARS-CoV-2 virus may have been identified much faster as part of the betacoronavirus genus being closely related to two bat-derived SARS-like coronaviruses (~88%), SARS-CoV (~79%) and MERS-CoV (~50%) species (6). This would not be possible in nearly all clinical laboratory tests available today including serology testing or multiplex PCR panels that are limited to a priori knowledge of selected pathogen targets. Furthermore, for well-classified species these (meta)genomes may be used to predict antimicrobial and antiviral resistance from known genetic determinants, however, may be limited by complex genetic relationships or transcriptional factors yet to be discovered (7). Numerous challenges still need to be addressed before this technology is adopted in routine practice, however, tremendous progress has been made in the scientific and regulatory front thus far discussed extensively in a review by Chiu CY et al. (8). The FDA has released general guidelines for the clinical validation of infectious disease testing, although there are no specific

recommendations for clinical implementation (9). However, a best practice approach should be taken including a failure mode and effects analysis using representative pathogens with continual assay evaluation and independent confirmation of unexpected results (8).

To date, several individual cases from nearly all types of clinical pathogens and samples have been reported to yield results from metagenomic sequencing where conventional tests were negative, often completely altering treatment and drastically improving outcome (8). One such case was a 14-year-old boy who developed hydrocephalus and status epilepticus necessitating a medically induced coma. The diagnostic workup including brain biopsy was unremarkable. However, metagenomic sequencing of cerebrospinal fluid revealed *Leptospira santarosai* infection, which was later confirmed by polymerase chain reaction tests, and treated with targeted antimicrobials which led to a full recovery and discharge a month later (10).

Failure of the current laboratory

The routine laboratory today currently fails to detect as much as 33-89% of respiratory pathogens, 40% of gastroenteritis and 60% of all encephalitis cases, which impede the surveillance of diseases (11–13). For instance, the Nipah viral outbreak in Malaysia in 1998 was falsely attributed to Japanese encephalitis, rendering early and appropriate control measures ineffective, which resulted in spread to other parts of Malaysia and nearby Singapore, which resulted in 105 deaths and the near collapse of a billion-dollar pig-farm industry (14). This viral outbreak which presents as encephalitis and respiratory illness has recently resurged since May 2018 in South India, with the last case reported in June 2019. Effectively our routinely laboratory capacity to detect viruses like Nipah virus has not fundamentally changed for decades. An effective treatment or vaccine for this disease does not exist, placing a greater need for early

outbreak detection and an aggressive response. Furthermore, given the high number of individual diagnostic tests requested for unidentified infections, a single and unbiased clinical metagenomic test may be less labor-intensive, quicker and overall more cost-effective.

A unique advantage is to obtain high-resolution transmission dynamics, capable of informing public health responses to terminate clusters of disease and avert similar events by understanding the origins of such outbreaks (15). Furthermore, understanding early on if this is truly an outbreak caused by a single strain or, simply an increased incidence involving multiple unrelated strains might prompt a faster and focused response that may include removal of the source, interrupting transmission and strengthening host defenses. Evidence of this was seen using portable sequencing device in the recent Usutu (16), Ebola (17), Zika (18), Lassa fever (19) and Yellow fever (20) outbreaks, a strategy termed as "precision public health". Other studies have recently combined genomic surveillance data with epidemiological data to identify transmission networks of SARS-CoV-2, which has identified cryptic transmission events including ward outbreaks of hospital-acquired infections and substantial transmission in healthcare associated community settings that had not been suspected by clinical or infection control teams, which may have implications for national public health policy (21, 22).

Current barriers to implementation

Numerous technical, clinical and regulatory challenges need to be addressed before metagenomic testing is widely available at the bedside. As various pathogens from bacteria and viruses to fungi and helminths may be found in multiple types of clinical samples, a single workflow is unlikely to suffice, but rather multiple workflows optimized to specific sample types may

emerge. Of particular importance for outbreak detection, a validated and standardized automated workflow to detect viruses from respiratory samples may have significant potential. However, technical challenges exists such as varying nucleic acid extraction efficacy, contamination issues and disproportionately high amounts of human DNA, which have been widely discussed elsewhere (8, 23, 24).

Although bioinformatic pipelines geared to detect known organisms have been well established, pipelines to detect novel organisms still need to be validated which might involve setting specific diagnostic thresholds and using techniques such as de novo assembly, a challenging feat with metagenomic samples (25). Furthermore, nucleic acid aligners may provide false-positive hits for non-human animal viruses and, may often require further analysis with higher stringency algorithms to improve specificity. While samples containing high levels of animal-related viruses may have numerous identified reads, more divergent pathogens or less optimal samples may require further testing and follow-up as seen in the discovery of Bas-Congo virus and a novel Astrovirus found in brain tissue which required the use of deep sequencing (26, 27).

Typically, investigations are initiated from a clinical perspective with a pre-determined hypothesis. This may be no different for metagenomics, as a general screening approach may be unfeasibly expensive, however, further studies are required to determine which patients and clinical syndromes are best suited for this test potentially including conditions such as fever of unknown origin, culture-negative endocarditis, aseptic meningitis, or when specific clusters of cases present with unknown etiology. Nevertheless, this non-targeted approach would require workstreams to follow up potentially significant findings, which has not been developed for general clinical laboratory use, and potentially refer this to central laboratories with clinical bioinformaticians for further assessment. Further research and follow-up studies are required to

distinguish among true findings, findings of unknown clinical significance and false positives from contamination or bioinformatic misclassification. If thresholds are too sensitive, announcement of pathogen detection proven to be untrue or not clinically significant could lead to loss of faith in the technology to detect emerging pathogens as in the discovery of xenotropic murine leukemia virus. This virus was erroneously linked to chronic fatigue syndrome that lead to some individuals starting antiretrovirals, although this was ultimately proven to be generated unintentionally in the laboratory through genetic recombination between two mouse retroviruses during propagation of a prostate-cancer cell line in the mid-1990s (28). Although other such challenges have been extensively discussed elsewhere (8, 29), progress is continually being made on these forefronts which may one day allow us to exploit metagenomics' full potential.

Timing, timing, timing

Time is of essence in an outbreak, with isolation delays in the source region and pathogen transmission risk largely determining the course of a potential pandemic (30). Although the delay in identification of SARS-CoV-2 from the earliest symptom onset to result was 37 days, in resource-limited settings, this timeframe can be prolonged even further where poor healthcare infrastructure exists with limited diagnostic capabilities (3). For instance, the West African Ebola outbreak was first identified on March 21, 2014, although the index case was traced as far back as nearly 3 months on Dec 26, 2013 (31). Automated metagenomic systems may offer resource-limited settings with the ability to improve diagnostics of complex infectious diseases and may also provide early warning signs, that would otherwise be missed due to lack of extensive testing infrastructure.

Disease X

Although the SARS-CoV-2 pandemic appears as an anomaly, it isn't. The last century also began with catastrophic waves of Spanish influenza that killed up to 100 million people globally (32). It is estimated that around one new disease emerges yearly (33), and although not all have pandemic potential, enough do (e.g Ebola virus, Middle East respiratory syndrome coronavirus) to pose a significant risk to global health and to task the World Health Organisation with an important mandate, which includes commissioning an expert committee to update its list of the most threatening infectious diseases that lack effective treatments or vaccines (34). However, what should be a warning signal to extensively fund research and technology for disease surveillance such as metagenomic technology is the entity at the end of this brief list-"Disease X". Last year's Disease X now has a name, COVID-19, while the next unknown Disease X is yet to come.

Beyond the healthcare setting

The development of systems to integrate metagenomic laboratory data with environmental samples such as sewage and animal samples from cattle and other livestock may likely yield insights into circulating agents relevant to population health. For instance, integrated surveillance networks will be useful to discern the relationship between antibiotic use and antibiotic resistance genes in receiving environments and clinical settings, offering specific opportunities for intervention. Furthermore, it has been recently demonstrated that wastewater testing may capture the increase and decrease of novel coronavirus cases in mid-sized metropolitan regions

prior to standard diagnostics, facilitating social distancing interventions or vaccine deployment (35). However, these systems need to be carefully developed protecting patient and institutional autonomy, while providing meaningful public health information. Further exploration is required to determine if these systems are based on public health need or if it should remain in the research domain.

The way forward

Clinical metagenomics is still a burgeoning field, and still requires considerable research and development before it becomes part of the clinical armamentarium. Numerous scientific challenges exist such as reducing high levels of human DNA in clinical samples, improved quality and contamination control, and workflow optimization and automation. Careful contextualization of results will be necessary before conclusions can be made, ideally within an agreed frame of standard operating procedures that is yet to be established. However, despite these challenges it is likely to be a routine test available to the physician in the future. This may emerge as an automated laboratory device, where minimal effort is required from laboratory personal. The report when ready would then be validated by a clinical microbiologist and made available to the treating physician. Potentially dangerous pathogens, and the closest genetic relative for unknown pathogens, should flag up immediately prompting further follow-up. This report and metagenomic data, together with other environmental and livestock data, may be synchronized to a regional or national surveillance hub, where imminent public health threats can be monitored, and a response initiated.

Thus far clinical metagenomics has been proven useful in detecting novel species and strains, outbreak transmission patterns and complicated microbiome diseases. Its role in the future of pandemic preparedness is anticipated and could exist as the earliest surveillance system we may have to detect outbreaks of unknown etiology, and to respond in an opportune manner. Despite the numerous challenges, clinical metagenomics' advantage of potentially becoming available in routine care for complicated infectious diseases offer the feasibility and scalability for added unbiased pathogen surveillance, potentially averting future outbreaks, and would be useful for both the individual patient and for the public health system.

Conflict of interest

The author declares no conflict of interest related to this work.

Author Bio

Dr. Kumeren Govender is a South African physician. He has received the Rhodes Scholarship and is based at the infectious disease unit at the University of Oxford. His research interest are clinical diagnostics, metagenomics and whole-genome sequencing.

References

1. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. 2008. Global trends in emerging infectious diseases. Nature 451:990–993.

2. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez A, Tomori O, Mazet JAK. 2018. The Global Virome Project. Science (80-) 359:872 LP – 874.

3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu

T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506.

4. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273.

5. Bharucha T, Oeser C, Balloux F, Brown JR, Carbo EC, Charlett A, Chiu CY, Claas ECJ, de Goffau MC, de Vries JJC. 2020. STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. Lancet Infect Dis.

6. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574.

7. Boolchandani M, D'Souza AW, Dantas G. 2019. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 20:356–370.

8. Chiu CY, Miller SA. 2019. Clinical metagenomics. Nat Rev Genet.

9. 2016. Food and Drug Administration. Infectious disease next generation sequencing based diagnostic devices: microbial identification and detection of antimicrobial resistance and virulence markers: draft guidance for industry and FDA. Rockville, MD.

10. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM,

Somasekar S, Federman S, Miller S. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med 370:2408–2417.

11. Ambrose HE, Granerod J, Clewley JP, Davies NWS, Keir G, Cunningham R, Zuckerman M, Mutton KJ, Ward KN, Ijaz S. 2011. Diagnostic strategy used to establish etiologies of encephalitis in a prospective cohort of patients in England. J Clin Microbiol 49:3576–3583.

12. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D. 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4.

13. Qi C, Hountras P, Pickens CO, Walter JM, Kruser JM, Singer BD, Seed P, Green SJ, Wunderink RG. 2019. Detection of respiratory pathogens in clinical samples using metagenomic shotgun sequencing. J Med Microbiol 68:996.

14. Looi LM, Chua KB. 2007. Lessons from the Nipah virus outbreak in Malaysia. Malays J Pathol 29:63–67.

15. Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9–20.

16. Munnink BBO, Münger E, Nieuwenhuijse DF, Kohl R, van der Linden A, Schapendonk CME, van der Jeugd H, Kik M, Rijks JM, Reusken C. 2020. Genomic monitoring to understand the emergence and spread of Usutu virus in the Netherlands, 2016–2018. Sci Rep 10:1–10.

17. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A. 2016. Real-time, portable genome sequencing for Ebola

surveillance. Nature 530:228–232.

18. Faria NR, da Silva Azevedo R do S, Kraemer MUG, Souza R, Cunha MS, Hill SC, Thézé J, Bonsall MB, Bowden TA, Rissanen I. 2016. Zika virus in the Americas: early epidemiological and genetic findings. Science (80-) 352:345–349.

19. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, Thielebein A, Hinzmann J, Oestereich L, Wozniak DM. 2019. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science (80-) 363:74–77.

20. Faria NR, Kraemer MUG, Hill SC, De Jesus JG, Aguiar RS, Iani FCM, Xavier J, Quick J, Du Plessis L, Dellicour S. 2018. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science (80-) 361:894–899.

21. Geoghegan JL, Ren X, Storey M, Hadfield J, Jelley L, Jefferies S, Sherwood J, Paine S, Huang S, Douglas J, Mendes FK, Sporle A, Baker MG, Murdoch DR, French N, Simpson CR, Welch D, Drummond AJ, Holmes EC, Duchêne S, de Ligt J. 2020. Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. Nat Commun 11:6351.

22. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S, Caller LG, Caddy SL. 2020. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis 20:1263–1272.

23. Dulanto Chiang A, Dekker JP. 2019. From the Pipeline to the Bedside: Advances and Challenges in Clinical Metagenomics. J Infect Dis 221:S331–S340.

24. Dekkera JP. 2018. Metagenomics for clinical infectious disease diagnostics steps closer to reality. J Clin Microbiol 56:e00850-18.

25. Ayling M, Clark MD, Leggett RM. 2020. New approaches for metagenome assembly with short reads. Brief Bioinform 21:584–594.

26. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, Sittler T, Veeraraghavan N, Ruby JG, Wang C, Makuwa M, Mulembakani P, Tesh RB, Mazet J, Rimoin AW, Taylor T, Schneider BS, Simmons G, Delwart E, Wolfe ND, Chiu CY, Leroy EM. 2012. A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa. PLOS Pathog 8:1–14.

27. Naccache SN, Peggs KS, Mattes FM, Phadke R, Garson JA, Grant P, Samayoa E, Federman S, Miller S, Lunn MP, Gant V, Chiu CY. 2015. Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult With Encephalitis by Unbiased Next-Generation Sequencing. Clin Infect Dis 60:919–923.

28. Paprotka T, Delviks-Frankenberry KA, Cingöz O, Martinez A, Kung H-J, Tepper CG, Hu W-S, Fivash MJ, Coffin JM, Pathak VK. 2011. Recombinant origin of the retrovirus XMRV. Science (80-) 333:97–101.

29. Greninger AL. 2018. The challenge of diagnostic metagenomics. Expert Rev Mol Diagn 18:605–615.

30. Caley P, Becker NG, Philp DJ. 2007. The waiting time for inter-country spread of pandemic influenza. PLoS One 2.

31. World Health Organization. 2014. Ground zero in Guinea: the Ebola outbreak smoulders –

undetected – for more than 3 months.

32. Iserson K V. 2020. The Next Pandemic: Prepare for “Disease X.” West J Emerg Med Integr Emerg Care with Popul Heal.

33. Knobler S, Mahmoud A, Lemon S, Mack A, Sivitz L, Oberholtzer K. 2004. Learning from SARS: preparing for the next disease outbreak.

34. Organization WH. 2020. Prioritizing diseases for research and development in emergency contexts. Geneva, Switzerland: World Health Organization.

35. Larsen DA, Wigginton KR. 2020. Tracking COVID-19 with wastewater. Nat Biotechnol 38:1151–1153.

Address for correspondence: Dr. Kumeren Govender, Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK; email: kumeren.govender@ndm.ox.ac.uk