Diagnostic Microbiology and Infectious Disease 84 (2016) 203–206

Contents lists available at ScienceDirect

Diagnostic Microbiology and Infectious Disease

journal homepage: www.elsevier.com/locate/diagmicrobio

Comparison of microarray-predicted closest genomes to sequencing for poliovirus vaccine strain similarity and influenza A phylogeny

Sebastian Maurer-Stroh [a,b,*], Charlie W.H. Lee [c], Champa Patel [d], Marilla Lucero [e], Hanna Nohynek [f], Wing-Kin Sung [c], Chrysanti Murad [g], Jianmin Ma [a], Martin L. Hibberd [c], Christopher W. Wong [c], Eric A.F. Simões [d,h]

[a] Bioinformatics Institute (BII), A*STAR, 30 Biopolis St, #07-01 Matrix, 138671, Singapore
[b] School of Biological Sciences, Nanyang Technological University (NTU), 60 Nanyang Drive, 637551, Singapore
[c] Genome Institute Singapore (GIS), A*STAR, 60 Biopolis St, #02-01 Genome, 138672, Singapore
[d] University of Colorado School of Medicine, 13001 E 17th Place, Aurora, CO 80045, USA
[e] Medical Department, Research Institute for Tropical Medicine, Alabang, Muntinlupa City, Philippines
[f] KTL National Public Health Institute, Helsinki, Finland
[g] Microbiology Department, Faculty of Medicine, Universitas Padjadjaran, Bandung, Indonesia
[h] Center for Global Health, Colorado School of Public Health, and Children's Hospital Colorado, 13001 E 17th Place, Aurora, CO 80045, USA

ARTICLE INFO

Article history:
Received 20 August 2015
Received in revised form 30 October 2015
Accepted 4 November 2015
Available online 6 November 2015

Keywords:
Infectious disease
Diagnostics
Epidemiology
PCR
Hybridization array
Poliovirus
Influenza

ABSTRACT

We evaluate sequence data from the PathChip high-density hybridization array for epidemiological interpretation of detected pathogens. For influenza A, we derive similar relative outbreak clustering in phylogenetic trees from PathChip-derived compared to classical Sanger-derived sequences. For a positive polio detection, recent infection could be excluded based on vaccine strain similarity.

1. Introduction

Using a random-tag PCR coupled to a high-density Affymetrix hybridization array, the PathChip, developed at the Genome Institute of Singapore, can detect tens of thousands of virus strains simultaneously (Wong et al., 2007). The sensitivity and specificity of the PathChip were comparable to other molecular diagnostic methods establishing its capability for use in routine diagnostic laboratories (Simões et al., 2013). We posited that the genomic data generated from use of this PathChip on respiratory specimens could be a useful adjunct in epidemiologic field studies and vaccine trials. Due to the nature of the PathChip design covering more than 25,000 viral genomes, the probe density for a single pathogen cannot be enough to derive exact sequences in the case of fast evolving or less frequently sampled pathogens, but instead, the closest hit of a sample to any sequence included on the chip can be calculated computationally. It is therefore important for proper interpretation to gauge the extent of differences of sequences derived from classical Sanger sequencing to the best PathChip hits for the respective same samples. A close match would suggest that potentially the pathogen chip could be used both for detection of viruses in samples and simultaneously provide data that could point to genetic relatedness of samples, paving the way for rapid molecular epidemiologic investigations.

2. Materials and methods

2.1. Philippine study population and respiratory sample collection

The samples for this study were obtained in Bohol, Philippines, between 2000 and 2004, from 12,194 participants aged 6 weeks to 23 months, with lower respiratory tract infections (LRTI), in a pneumococcal vaccine trial, as detailed before (Lucero et al., 2009). In parallel, older children from the 6 study areas, aged 24–59 months, not in the trial, with LRTI were enrolled in a separate epidemiologic study. Informed consent was obtained from parents or guardians of all children

* Corresponding author. Tel.: +65-6478-8377; fax: +65-6478-9047.
E-mail address: sebastianms@bii.a-star.edu.sg (S. Maurer-Stroh).

http://dx.doi.org/10.1016/j.diagmicrobio.2015.11.003

for participation in the vaccine trial, and separate consent was obtained for each child to collect respiratory tract specimens. The institutional boards (IRBs) at the RITM in Manila, Philippines; KTL Finland; and COMIRB, Aurora, USA, approved the study.

In total, 2066 nasopharyngeal aspirate samples were obtained from these subjects. Batches of specimens were transported to Denver, CO, USA, on dry ice, where they were stored at −86 °C. From these, 290 samples were selected for a validation study of the PathGEN PathChip platform in comparison with other pathogen diagnostics systems to determine the robustness of this platform. The wet lab processing methods used are described in the previous work (Simões et al., 2013). This current work adds the epidemiological detailed perspective for 14 influenza A H3N2–positive samples and 1 poliovirus-positive sample for which Sanger-derived sequences could be obtained (see below).

2.2. Influenza samples

The nasal wash samples (Lucero et al., 2009) which were previously identified as having influenza (Simões et al., 2013) were assayed using PCR to subtype and sequence the sample. RNA was extracted from the nasal washes using TRIzol LS (Life Technologies, Grand Island, NY, USA) according to the manufacturer's suggested protocol. cDNA for influenza was produced using SuperScript III first strand synthesis system for RT-PCR (Life Technologies) using the primer Uni12W for influenza A samples (WHO, 2009). PCR for the hemagglutinin was completed using a nested PCR approach. World Health Organization primers and protocols were used to obtain the genes and then sequence them (WHO, 2009). The bands were visualized on a 1% agarose gel then sequenced using the Sanger method.

2.3. Poliovirus sample

A nasal wash sample (Lucero et al., 2009) which was identified as possibly being positive for poliovirus (Simões et al., 2013) underwent PCR for the VP1 region. The VP1 region was selected for PCR and sequencing for its specific defining properties of poliovirus (Oberste et al., 1999). cDNA was produced using the SuperScript III first-strand synthesis system for RT-PCR kit (Life Technologies) using random hexamers. PCR was carried out using the previously described Y7R and Q8 primers (Kilpatrick et al., 2011) using Thermo-Start polymerase (Thermo Scientific, Pittsburgh, PA, USA) according to the manufacturer's recommendations. The cycling conditions were as follows: 95 °C for 15 minutes, 35 cycles of 95 °C for 30 seconds, 52 °C for 30 seconds, and 72 °C for 1 minute, followed by 72 °C for 7 minutes. The band was visualized on a 1% agarose gel then sequenced using the Sanger method.

2.4. Indonesian study population and respiratory sample collection

The Indonesian samples were obtained from a population-based study in rural West Java, Bandung, Indonesia, to estimate the burden of influenza in 2008. We conducted daily surveillance in the first level health center (Puskesmas) in Soreang and Cileunyi subdistricts, Bandung West Java Province, Indonesia. A total of 197 children and adults with Influenza-like illness (ILI) were recruited (nasal wash/swab specimens), and PCR of nasal specimens revealed that 69 had influenza A and 1 had influenza B. The PCR protocol is identical to the previous study, and IRB approval is included as listed above with additional IRB approval at RS Hasan Sadikin hospital, Indonesia. Four samples that were confirmed to be H3N2 influenza A in 2008 were also tested with the PathChip, and the closest matching sequence was determined (see below).

2.5. PathChip and PathChip-derived sequences

The PathChip and its usage protocol have been extensively described before (Simões et al., 2013; Wong et al., 2007). Briefly, the basic assumption underlying the PathChip protocol assumes that the researcher/physician does not know what pathogen is contained within the sample. As such, sample preparation includes a single-tube PCR reaction, using a semirandom approach. PCR bias is avoided during the assay design process, using the LOMA algorithm we published previously (Lee et al., 2008). The next step is hybridization to an array in Affymetrix custom GeneChip cartridge format consisting of probes uniformly spaced out across 25,000 full viral genome sequences from 59 genera organized in 154 clinically relevant groups. The PDA algorithm was used to predict the pathogen recognition signatures for each genome on the chip (Wong et al., 2007). For influenza A, B, and C and poliovirus, the chip provides genome coverage for 13467, 2673, 204, and 56 strains/isolates, respectively. Relevant for this study, each H- or N- influenza subtype is detected by separate pathogen recognition signatures comprising an average of 100 probes. The Affymetrix image file is automatically analyzed by the accompanying software (Wong et al., 2007) which presents as output assay quality control metrics and the identified pathogen recognition signatures, including the NCBI GenInfo number of the closest matching pathogen genome in the database (Simões et al., 2013). In this study, we refer to a PathChip-derived sequence as the sequence of the closest known genome hit detected by PathChip.

2.6. Multiple alignments and phylogenetic trees

For each nucleotide sequence set derived from the 2 methods, multiple alignments were created with MAFFT using L-INS-I parameters (Katoh and Standley, 2014), and phylogenetic trees were inferred by using the maximum likelihood method based on the Tamura–Nei model (Tamura and Nei, 1993) with a discrete gamma distribution to consider evolutionary rate differences among sites (5 categories, including invariable) in MEGA6 (Tamura et al., 2013). For Fig. 1D, influenza A H3N2 vaccine strains were downloaded from the NCBI Influenza Virus resource (Bao et al., 2008) and multiple alignments and phylogenetic trees derived as described above.

2.7. Tree similarity

To test significance of phylogenetic tree similarity, we compare the measure of exactly agreeing triplets (Estabrook et al., 1985) as distance between the Sanger and PathChip hit trees with the same distance for randomly permutated trees (randomization tests over 100 trees as implemented in TOPD (Puigbò et al., 2007)).

3. Results and discussion

First, we evaluated a hit for a poliovirus identified in 1 of the specimens from the previous study of samples from the Philippines (Simões et al., 2013). Compared to classical Sanger sequencing, the PathChip-derived best hit sequence was found to be 96% identical over a sequence length of 1000 nucleotides (VP1 region). While this is clearly limited by the sampling density of polioviruses available in the databases, both the Sanger- and PathChip-derived sequences were more closely related to the Sabin type 1 vaccine reference strain (98.5% and 97.5%, respectively) than to recent Philippine outbreak strains (95.6% and 95.8%) suggesting recent immunization for this patient with the live attenuated vaccine and not infection with a wild-type strain circulating at the time (Shimizu et al., 2004).

Next, we have previously shown that the PathChip can readily distinguish different influenza subtypes with hits for influenza A H3N2 and H1N1 as well as influenza B and the rare or less well studied influenza C. In this study, we attempt the more difficult task of within subtype clustering. We had to focus on H3N2 since the numbers of detection of the other subtypes were too few to be interpreted epidemiologically (Simões et al., 2013).

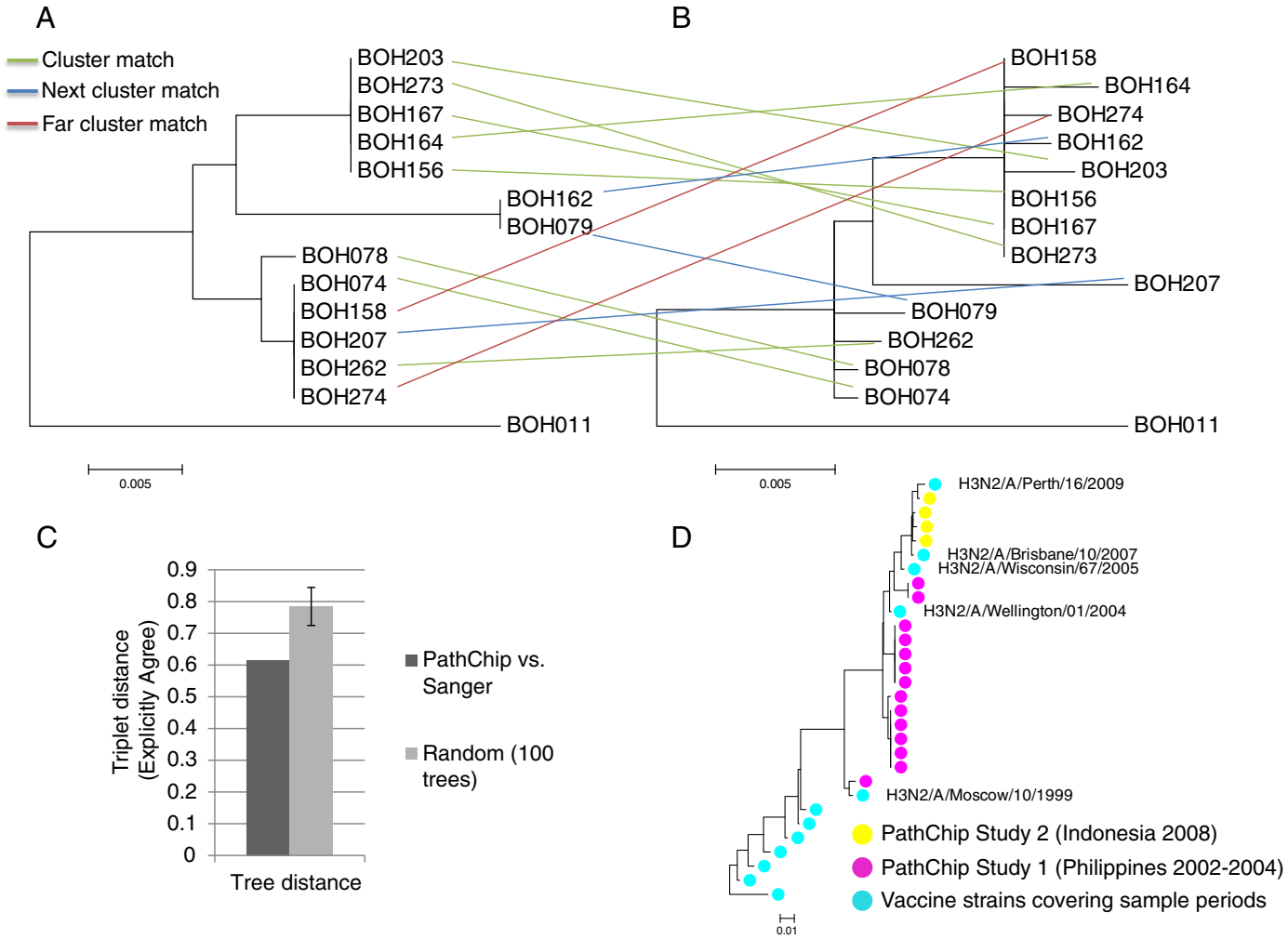

Fig. 1. PathChip hit trees match Sanger sequence trees. H3 hemagglutinin phylogenetic trees of (A) PathChip hits and (B) Sanger sequences both derived from the same samples. Colored bars connect corresponding samples. (C) Tree distance measure for the 2 trees above compared to average distance of 100 randomly permutated trees. (D) Phylogenetic tree of PathChip hit sequences with H3N2 vaccine strains.

Fourteen samples could be sequenced directly from clinical material using classical Sanger protocols, and sequences were also computationally derived from the PathChip best hits result for the same samples. On average, the PathChip-derived sequences were 98.9% identical to their Sanger controls (97.6–99.7%) over sequence lengths from 1200 to 1700 nucleotides, which is substantially better than for the poliovirus example due to the intense global influenza surveillance efforts. Nevertheless, BLAST (Altschul et al., 1997) hits of the Sanger-derived sequences against GenBank (Benson et al., 2011) indicate that they are most similar to clusters in Malaysia, Hong Kong, or Australia, respectively, and not the Philippines showing that influenza A H3N2 strains in the Philippines during the sample period were not well represented in the database. This also means that the immediate virus strains from the obtained samples are also not included in the used version of the chip, therefore representing a realistic scenario of how the PathChip hit profile result would look like for slightly distinct or drifted strains, expected during typical influenza A evolution (Rambaut et al., 2008). Interestingly, the influenza samples analyzed with PathChip could nevertheless be distinguished among each other by their relative hit profile to different groups of strains reminiscent of outbreak clustering seen with normal sequences.

To see if this relative clustering matches that of Sanger-derived sequences, we created phylogenetic trees for each nucleotide sequence set derived from the 2 methods. Fig. 1A and B shows that the corresponding phylogenetic trees for both methods are, as expected, not perfect but indeed similar. Nine of the 14 samples are exactly matched to the same cluster in the trees, and another 3 match at least to the neighboring cluster asserting reasonable agreement in 86% of the samples. To test significance of this tree similarity, we compare the Sanger and PathChip hit trees with randomly permutated trees and find that the agreement of PathChip hit and direct sequence trees is at least 2 SDs better than the random average (Fig. 1C). Agreement of triplets (measure used here, see Methods) is important for interpretation of samples with regard to outbreak clustering indicating that PathChip results could be used for this.

Another validation is if the identified best PathChip hit sequences would fit into the known evolutionary pattern and history of influenza as reflected by WHO-recommended vaccine strains. Indeed, all sample sequences are found between or with vaccine strains of their expected time period 2002–2004 (Fig. 1D, pink dots). To show that the approach is robust and able to distinguish samples from different seasons, we applied the same procedure for H3N2 influenza A samples from another surveillance study in Indonesia in 2008 where PCR-positive hits were also concurrently tested with the PathChip. Also in this case, the PathChip-derived sequences are fitting to their expected time period in between the respective vaccine reference strains (Fig. 1D, yellow dots).

We are aware that the necessary array coverage to obtain reasonable matches through closest genome hit matches strongly depends on the respective pathogen's representation in the database used to design

the array/chip. We chose here influenza as example because there is good database coverage due to the global surveillance efforts, and it is an important commonly encountered pathogen where epidemiological interpretation can help in public health decisions. Given that there remains a discrepancy of the closest genome match with the Sanger-derived sequences, secondary confirmation of the sequence details would be recommended by a complementary method, while our and similar chip protocols could be useful for large-scale screens to identify which samples and sequences are of greater interest to be studied in more detail.

4. Conclusions

In summary, we show that PathChip-derived best hit sequences can be used for epidemiological interpretation through relative phylogenetic tree clustering even for fast evolving pathogens such as influenza. Specifically, one can identify outbreak clusters including vicinity to prevailing vaccine strains. We note that these findings on usability of relative hit clustering may not be unique to the tested PathChip platform but could be equally useful for other PCR-coupled hybridization array systems for pathogen detection, although this requires further studies.

Conflicts of interest

Eric A. F. Simões, Champa Patel, Marilla Lucero, Hanna Nohynek, Chrysanti Murad, Jianmin Ma, and Sebastian Maurer-Stroh declare that they have no conflict of interest. Four authors (Christopher W. Wong, Wing-Kin Sung, Martin L. Hibberd, and Charlie W. H. Lee) are co-founders and directors of PathGEN Dx Pte, Ltd, which has licensed the PathChip technology from A*STAR.

Acknowledgments

This study is part of the research of the Acute Respiratory Infection Vaccine (ARIVAC) consortium. We are indebted to the consortium study team and the following collaborators: The Data Safety Monitoring Board: Kim Mulholland (chair), Keith Klugman, Mary Ann Lansang (local safety monitor), David Sack, Pratap Singhashivanon, Peter Smith, and Chongsuphajaisiddhi Tan; Research Institute for Tropical Medicine (RITM): Socorro Lupisan, Beatriz Quimbao, Diozele Sanvictores, Erma Abucejo-Ladesma, Juanita Ugpo, Marites Lechago, Leilani T. Nillos, Vernoni Ermata Dulalia; National Institute of Health and Welfare (formerly National Institute of Public Health KTL): Taneli Puumalainen, Antti Nissinen, Anu Soininen, Petri Ruutu, and P. Helen Mäkelä; and University of Colorado: Shirstine Gorton and Adriana Weinberg; University of Queensland: Ian Riley, Margaret de Campo, Sanofi Pasteur, Emmanuel Feroldi, Dirk Teuwen, and James Maleckar.

The ARIVAC consortium thanks and acknowledges the participation of the infants, parents, staff, Local Government of the Province of Bohol and local government units of Baclayon, Balilihan, Cortes, Dauis, Panglao, and Tagbilaran City; staff of the pathology and pediatric departments of the Bohol Regional Hospital, and the private hospitals Tagbilaran Community Hospital, Borja Family Clinic, Medical Mission Group of Hospitals, Ramiro Community Hospital, St Jude Hospital, Englewood Hospital, and Tagbilaran Puericulture Center.

Support for research to enable this publication was provided by the European Commission DG Research INCO program (contracts IC18-CY97-2019, ICA4-CT-1999-10008, and ICA4-CT-2002-10062), Academy of Finland (contracts 206283, 106974, 108873, and 108878), Finnish Ministry of Foreign Affairs (bilateral contracts 75502901 and 327/412/2000), Finnish Physicians for Social Responsibility, GAVI ADIP Pneumo, Sanofi Pasteur, Research Institute for Tropical Medicine of the Philippines, National Public Health Institute Finland, University of Queensland, University of Colorado, National Health and Medical Research Council of Australia, PATH and the Agency for Science, Technology and Research (A*STAR) Singapore (grant IAF311010).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402.

World Health Organization (WHO). WHO information for laboratory diagnosis of pandemic (H1N1) 2009 virus in humans—revised. Accessed March 11, 2015. Available at: http://www.who.int/csr/resources/publications/swineflu/WHO_Diagnostic_RecommendationsH1N1_20090521.pdf, 2009.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. J Virol 2008;82(2):596–601.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2011;39(Database issue):D32–7.

Estabrook GF, McMorris FR, Meacham CA. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Biol 1985;34(2):193–200.

Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol Biol 2014;1079:131–46.

Kilpatrick DR, Iber JC, Chen Q, Ching K, Yang S-J, De L, et al. Poliovirus serotype-specific VP1 sequencing primers. J Virol Methods 2011;174(1–2):128–30.

Lee WH, Wong CW, Leong WY, Miller LD, Sung WK. LOMA: a fast method to generate efficient tagged-random primers despite amplification bias of random PCR on pathogens. BMC Bioinformatics 2008;9:368.

Lucero MG, Nohynek H, Williams G, Tallo V, Simões EAF, Lupisan S, et al. Efficacy of an 11-valent pneumococcal conjugate vaccine against radiologically confirmed pneumonia among children less than 2 years of age in the Philippines: a randomized, double-blind, placebo-controlled trial. Pediatr Infect Dis J 2009;28(6):455–62.

Oberste MS, Maher K, Kilpatrick DR, Pallansch MA. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. J Virol 1999;73(3):1941–8.

Puigbò P, Garcia-Vallvé S, McInerney JO. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 2007;23(12):1556–8.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature 2008;453(7195):615–9.

Shimizu H, Thorley B, Paladin FJ, Brussen KA, Stambos V, Yuen L, et al. Circulation of type 1 vaccine-derived poliovirus in the Philippines in 2001. J Virol 2004;78(24):13512–21.

Simões EAF, Patel C, Sung W-K, Lee CWH, Loh KH, Lucero M, et al. Pathogen chip for respiratory tract infections. J Clin Microbiol 2013;51(3):945–53.

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10(3):512–26.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013;30(12):2725–9.

Wong CW, Heng CLW, Wan Yee L, Soh SWL, Kartasasmita CB, Simoes EAF, et al. Optimization and clinical validation of a pathogen detection microarray. Genome Biol 2007;8(5):R93.