Title	Application of large-scale sequencing and data analysis to research on emerging infectious diseases
Author(s)	Liu, Y; Lam, TY; Zhu, H; Guan, Y
Citation	21st International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology (VEME), Seoul, Korea, 14-19 August 2016. In Virus Evolution, 2017, v. 3 n. Suppl 1, p. vew036.023
Issued Date	2017
URL	http://hdl.handle.net/10722/247081
Rights	Pre-print: Journal Title] ©: [year] [owner as specified on the article] Published by Oxford University Press [on behalf of xxxxxx]. All rights reserved. Pre-print (Once an article is published, preprint notice should be amended to): This is an electronic version of an article published in [include the complete citation information for the final version of the Article as published in the print edition of the Journal.] Post-print: This is a pre-copy-editing, author-produced PDF of an article accepted for publication in [insert journal title] following peer review. The definitive publisher-authenticated version [insert complete citation information here] is available online at: xxxxxxx [insert URL that the author will receive upon publication here].; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

study of HIV-1 transmitted drug resistance, filled questionnaires were obtained for 440 patients. Homosexual contact as the most probable mode of HIV acquisition reported 326/440 patients. Subsequently, partial *pol* sequences were obtained for 252 MSM patients who were included in the present study. Sequences were analyzed using the following automatic subtyping tools: REGA 2.0, REGA 3.0, COMET HIV-1 1.0, jpHMM, and SCUEAL. Sequences that gave divergent subtyping results were considered to be potential recombinant forms. Temporal trend of the proportion of non-B and potential recombinant sequences (both combined termed as "non-pure subtype B" sequences) was evaluated with Fisher exact test used for the assessment of statistical significance. All five subtyping tools gave concordant subtyping results in 230/252 (91.3%) of sequences. Pure subtype B was assigned to 226/252 (89.7%) sequences and subtype A, subtype C, subtype F and CRF01_AE were determined in one patient each (0.4%). The remaining 22/252 (8.7%) sequences yielded divergent results with at least one of the subtyping tools, an indication to a possible recombination event in the past. An increase in the proportion of "non-pure subtype B" HIV-1 variants was noted over the years with 0% (95% confidence interval (CI): 0–16%), 5% (95% CI: 1–15%), 4% (95% CI: 0–13%), 11% (95% CI: 4–21%), and 25% (95% CI: 15–39%) determined in the years 2000–2, 2003–5, 2006–8, 2009–11, and 2012–4, respectively. The marked increase was on account of an increasing number of potential recombinant sequences with subtype B as one of the founder subtypes, since this subtype was identified in 21/22 of divergent sequences. The remaining divergent sequence was a complex recombinant containing subtypes D and G. Additionally, all 4 pure non-B subtyped sequences were determined in patients diagnosed in the last two years (2013–4). The results obtained indicate that subtypes other than B are entering the HIV-1 MSM epidemic in Slovenia, making recombination events among different subtypes more plausible. A marked increase in the numbers of non-B subtypes and potential recombinant forms was observed among MSM in Slovenia in recent years.

A23 Identification of HIV drug resistance mutation patterns using illumina MiSeq next generation sequencing in patients failing second-line boosted protease inhibitor therapy in Nigeria

Onyemata E. James,[1] Ledwaba Johanna,[3] Datir Rawlings,[1] M. Babajehson,[1] Ismail Ashrad,[3] Derache Anne,[4] Alash'le Abimiku,[1,2] Patrick Dakum,[1] Hunt Gillian,[3] Nicaise Ndembi,[1]

[1]Institute of Human Virology - Nigeria, Abuja, Federal Capital Territory, Nigeria, [2]Institute of Human Virology, University of Maryland School of Medicine, MD, USA, [3]AIDS Research Unit, National Institute for Communicable Diseases, National Health Laboratory Services, Johannesburg, South Africa and [4]Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa

There is limited information of patterns of protease inhibitor (PI) resistance in adults failing second-line therapy. Next generation sequencing (NGS) detects drug resistance mutations as low as 1%. However, the clinical implications of these minority variants to treatment outcomes are still in debate. Approximately 5% of antiretroviral treatment (ART) exposed patients in our treatment program are on second-line boosted protease inhibitor (PI). Population-based sequencing conducted on some of these patients revealed no major HIV drug resistance mutations (DRMs) to PIs. We compared population-based sequencing and NGS results with the view of identifying patterns of drug resistance mutations and minority variants. Forty-eight plasma samples from 40 patients on second-line NRTI/NNRTI and boosted PI regimens with evidence of virologic failures (VL ≥1,000 copies/ml) were used in this study. Of these, eight were obtained at the time of first-line failure while the remaining 40 at the time of second-line failure. Ultra-deep sequencing sample preparation was achieved using Illumina Nextera XT protocol. This required that target amplicon be subjected to fragmentation, tagging, indexing, size exclusion bead purification, normalization, and pooling. MiSeq data analysis was performed using the Geneious software by applying 1% cut-off at major drug resistance sites. Electropherogram data were generated using ABI 3130 genetic analyzer and analysis performed using Stanford Genotyping Resistance Interpretation Algorithm available at http://sierra2.stanford.edu/sierra/servlet/JSierra and IAS-USA 2015 Drug Resistance Interpretation list. MiSeq sequencing showed that 53% ($n=23$) of the patients developed PI resistance, 93% ($n=40$) had NRTI resistance, and 70% ($n=30$) had NNRTI resistance. Of the DRMs detected in protease, L90M mutation was the most common mutation (28%, $n=12$) followed by L76V (21%, $n=9$), then I47V (7%, $n=3$) and I84V (7%, $n=3$). Among the NRTI associated mutations L74V was the predominant mutation (77%, $n=33$) followed by M184V/I (60%, $n=26$) then TAMS (51%, $n=22$). Of these, 33% of patients ($n=14$) showed NRTI + NNRTI mutations, 39% ($n=17$) showed NRTI + NNRTI + PI mutations, 7% ($n=3$) showed NRTI + PI mutations, whereas 21% ($n=9$) and 2.3% ($n=1$) exclusively showed NRTI and NNRTIs mutations respectively. Twenty-eight samples that had both MiSeq and Sanger sequencing data were available for a comparison of mutational patterns in the PI region. Miseq sequencing revealed minority PI mutations in 10 samples that were wild type by Sanger sequencing and one sample showed mutations in both Sanger and NGS. The ten samples revealing mutations based on MiSeq data comprised of minority variants including L90M (50%, $n=5$), L76V (20%, $n=2$), I50V (10%, $n=1$), and N88S (20%, $n=2$). Our data suggest that even in the absence of PI mutations based on Sanger data, those minority variants can be present. NGS revealed the presence of PI resistance mutations in patients who had wild-type using population-based sequencing. Given that patient regimen revealed that minority variants were unlikely selected by ART pressure, our results suggest poor adherence as the likely contributor to second-line failure due to the high genetic barrier of PIs. Since ART adherence in these patients was monitored using clinico-immunological parameters and virological tests only when treatment failure was suspected, our results suggest the need for routine virological monitoring. This should provide early opportunity for adherence intervention and thereby avoiding the need for switch to salvage or third-line treatment options, which is more expensive and not readily available in our setting.

A24 Application of large-scale sequencing and data analysis to research on emerging infectious diseases

Yongmei Liu, Tommy Tsan-Yuk Lam, Huachen Zhu, and Yi Guan

State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China

Many human diseases are caused by emerging pathogens, such as the SARS and MERS coronaviruses. Timely understanding of the behaviors of these pathogens plays an important role in helping doctors and scientists in searching for treatment methods and designing vaccines. The development of next-generation sequencing (NGS) has led to significant breakthroughs in the production of large amount of unbiased DNA sequence data from field and human clinical samples, providing the capacity

to identify the sources of infection, and the virus evolution as well as host/virus interaction. In our study, using 454/Illumina sequencing, we have obtained large amount of whole genome sequences. We designed a preliminary bioinformatics analysis pipeline to classify these NGS reads. First we mapped our nucleotide reads to GenBank reference sequences using BLAST, and classified them by their taxonomic family, such as host, virus and unclassified. Then, for a specific type of virus (e.g. influenza virus, MERS coronavirus), we conducted de novo and reference based assembly of the reads to obtain the full genome sequences for further phylogenetic study. In the future, through advanced bioinformatics tools, we hope to get more detailed information from our large amount of NGS sequences of field/clinical samples, experimental data, especially in the following areas: (i) finding novel pathogens in unclassified sequences; (ii) virus/virus interactions; (iii) pathogen/host interaction.

A25 Phylogenetic analysis of the nucleocapsid and RNA-dependent RNA polymerase fragments of the first imported case of middle east respiratory syndrome coronavirus (MERS-CoV) infection in the Philippines from Saudi Arabia, February 2015

I.A.P. Medado,[1,*] J.R.C. Orbina,[1] N.T.M. Yabut,[1] L.L.M. Dancel,[1] T.C. Tan,[1] M.A.U. Igoy,[1] A.M.R. Mojica,[1] I.C. Lirio,[1] A.C. Ablola,[1] B.C. Mateo,[1] M.A.J. Biol,[1] A.D. Nicolasora,[1] H.L.E. Morito,[1] K.I.M. Cruz,[1] C.M.F. Roldan,[1] P.B. Medina,[2] E.S. Mercado,[1] C.S. Demetria,[2] R.J. Capistrano,[3] S.P. Lupisan [4]

[1]Molecular Biology Laboratory, Research Institute for Tropical Medicine - Department of Health, Philippines, [2]Special Pathogens Laboratory, Research Institute for Tropical Medicine - Department of Health, Philippines, [3]Surveillance and Response Unit, Research Institute for Tropical Medicine - Department of Health, Philippines and [4]Research Institute for Tropical Medicine - Department of Health, Philippines

We report the first laboratory-confirmed case of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) infection from a patient returning to the Philippines from the Kingdom of Saudi Arabia (KSA). MERS-CoV was first identified in 2012 circulating in Middle Eastern countries with outbreaks occurring in KSA, the United Arab Emirates (UAE), and South Korea, plus sporadic imported cases in at least 20 other countries. The Philippines is at risk for MERS-CoV transmission from frequent travelers, such as overseas Filipino workers and Hajj pilgrims, coming from Middle Eastern countries. Throat swabs, sputum samples, and a rectal swab were collected from the index case within 13 to 22 days after the onset of symptoms. MERS-CoV testing was performed using a real-time reverse transcription polymerase chain reaction (RT-qPCR) screening assay targeting regions upstream of the envelope gene (upE) and the nucleocapsid gene (N2), a confirmatory RT-qPCR assay targeting regions within the open reading frame 1a gene (ORF1a) and another region of the N gene (N3), and Sanger sequencing of regions of the N and RNA-dependent polymerase (RdRp) genes. The index case tested weakly positive for MERS-CoV in a sputum sample until day 19 of illness. Sequences of the N and RdRp gene regions reveal 100 and 99% similarity with MERS-CoV sequences obtained in KSA and UAE, respectively, confirming that the infection originated from Middle Eastern strains. Two unique synonymous/silent mutations (T15259A and T15265C) were identified in the RdRp sequence fragments. Whole genome sequencing of the strain may identify other mutations across the genome and determine the most probable origin of the strain.

A26 Transmission patterns and evolution of RSV in a community outbreak identified by genomic analysis

Charles N. Agoti,[1,2] Patrick K. Munywoki,[1,2], My V. T. Phan,[3] James R. Otieno,[1] Everlyn Kamau,[1] Anne Bett,[1] Ivy Kombe,[1] George Githinji,[1] Graham F. Medley,[4] Patricia A. Cane,[5] Paul Kellam,[3,6] Matthew Cotton,[3] D. James Nokes,[1,7]

[1]Epidemiology and Demography Department, KEMRI – Wellcome Trust Research Collaborative Programme, Kilifi, Kenya, [2]School of Health and Human Sciences, Pwani University, Kilifi, Kenya, [3]The Wellcome Trust Sanger Institute, Cambridge, UK, [4]Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK, [5]Public Health England, Salisbury, UK, [6]Imperial College, London, UK and [7]School of Life Sciences and WIDER, University of Warwick, Coventry, UK

Detailed information on the source, spread and evolution of respiratory syncytial virus (RSV) during seasonal community outbreaks remains sparse. Molecular analyses of attachment (G) gene sequences from hospitalised cases suggest that multiple genotypes and variants co-circulate during epidemics and that RSV persistence over successive seasons is characterized by replacement and multiple new introductions of variants. No studies have defined the patterns of introduction, spread and evolution of RSV at the local community and household level. We present a whole genome sequence analysis of 131 RSV group A viruses collected during six-month household-based RSV infection surveillance in Coastal Kenya, 2010 within an area of 12 km^2. RSV infections were identified by regularly screening of all household members twice weekly. Phylogenetic analysis revealed that the RSV A viruses in 9 households were closely related to genotype GA2 and fell within a single branch on the global phylogeny. Genomic analysis allowed the detection of household-specific variation in seven households. For comparison, using only G gene analysis, household-specific variation was found only in 1 of the 9 households. Nucleotide changes were observed intra-host (viruses identified from same individual in follow-up sampling) and inter-host (viruses identified from different household members) and these coupled with sampling dates enabled partial reconstruction of the within household transmission chains. The genomic evolutionary rate for the household dataset was estimated as 2.307×10^{-3} (95% highest posterior density: $0.93513–4.1636 \times 10^{-3}$) substitutions/site/year. We conclude that (i) at the household level, most RSV infections arise from the introduction of a single virus variant followed by accumulation of household specific variants and (ii) analysis of complete virus genomes is crucial to better understand viral transmission in the community. A key question arising is whether prevention of RSV introduction or spread within the household by vaccinating key household members in these functions would lead to a reduced onward community wide transmission.

A27 Using whole genome sequence data and minority variant profiles to elucidate transmission patterns during RSV household outbreaks

George Githinji,[1] Charles Agoti,[1,2] Patrick Munywoki,[1,2] Anne Bett,[1] Paul Kellam,[3] Matt cotton,[3] D. James Nokes,[1,4]

[1]KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya, [2]Pwani University, Kilifi, Kenya, [3]Wellcome Trust Sanger Institute, Hinxton, UK and [4]School of Life Sciences and WIDER, University of Warwick, UK

Reconstructing transmission chains for outbreaks is important in understanding how viruses spread. Furthermore, defining the main underlying determinants of transmission chains is important for developing effective interventions. Whole