Deliverable

D6.1. Report on next-generation sequencing (NGS) based diagnostics in comparison to gold standard methods

Dissemination level: PU

Version: 1.0

Due: Month 55

Completed: Month 58

Authors: Victoria Janes (AMC), Constance Schultsz (AMC), Sander van Boheemen (EMC), Guido Werner and Stephan Fuchs (RKI)

Contributing partners: AMC, EMC, RKI

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

Deliverable Description

Work-package 6 (WP6) aims to utilize state-of-the-art tools, technologies, and methods developed in WP3 for the generation of comparable data and turning these data into actionable information for clinical diagnostic use and hospital epidemiology. The acceptance of next generation sequencing (NGS) in the clinical laboratory relies heavily on the performance criteria relevant to clinicians, such as speed, cost, positive and negative predictive values, and relevance for treatment guidance. Therefore, a comparison between conventionally used gold-standard methods in clinical diagnostic laboratories and NGS is needed to evaluate the performance of NGS and its advancements to better guide treatment.
In this deliverable, we reported the comparison of NGS versus gold-standard methods in different steps of the diagnostic pipelines, from sample preparation, analytical workflow, to application of NGS in outbreak investigation. These comparisons were reported in three reports with two publications and one pilot study:

1. Report on the comparison in the performance of Whole Genome Sequencing versus Amplified Fragment Length Polymorphism in the management and characterization of a 5-month nosocomial Vancomycin-Resistant Enterococcus outbreak (AMC)
2. Report on the comparison of NGS (i.e. metagenomics sequencing and viral metagenomics sequencing using virus-specific baits-ViroCap) versus gold standard methods (i.e. PCR amplification of 10 overlapping amplicons and sequencing) for investigating the complete genome of HIV from plasma samples
3. Report on the performance of Biomerieux versus Seqsphere for genotyping based on WGS data of ~300 *Staphylococcus aureus* strains in comparison with traditional typing methods (MLST, spa-typing using PCR and Sanger sequencing, antibiotic susceptibility testing using broth microdilution)

Reports

1. Report on the applications of Whole Genome Sequencing and Amplified Fragment Length Polymorphism in the management and characterization of a nosocomial Vancomycin-Resistant Enterococcus outbreak (AMC)

Introduction

As unrelated, but genetically similar antimicrobial-resistant (AMR) pathogens may circulate simultaneously, rapid high-resolution molecular typing methods are needed for outbreak management. We compared Amplified Fragment Length Polymorphism and WGS typing for the management and characterization of a nosocomial outbreak.

Publication

The output of this study is a publication:

Janes VA, Notermans DW, Spijkerman IJB, Visser CE, Jakobs ME, van Houdt R, Willems RJL, de Jong MD, Schultsz C, Matamoros S. *Amplified fragment length polymorphism and whole genome sequencing: a comparison of methods in the investigation of a nosocomial outbreak with vancomycin resistant enterococci*. Antimicrob Resist Infect Control. 2019 Sep 23; 8:153. doi: 10.1186/s13756-019-0604-5. PMID: 31572571; PMCID: PMC6757385.

The following is the abstract of the manuscript.

Abstract

Amplified Fragment Length Polymorphism and Whole Genome Sequencing: a comparison of methods in the investigation of a nosocomial outbreak with Vancomycin-Resistant Enterococci

Background. Recognition of nosocomial outbreaks with antimicrobial-resistant (AMR) pathogens and appropriate infection prevention measures are essential to limit the consequences of AMR pathogens to patients in hospitals. Because unrelated, but genetically similar AMR pathogens may circulate simultaneously, rapid high-resolution molecular typing methods are needed for outbreak management. We compared amplified fragment length polymorphism (AFLP) and whole-genome sequencing (WGS) during a nosocomial outbreak of vancomycin-resistant *Enterococcus faecium* (VRE) that spanned five months.

Methods. Hierarchical clustering of AFLP profiles was performed using unweighted pair-grouping, and similarity coefficients were calculated with Pearson correlation. For WGS-analysis, core single nucleotide polymorphisms (SNPs) were used to calculate the pairwise distance between isolates, construct a maximum likelihood phylogeny and establish a cut-off for the relatedness of epidemiologically linked VRE isolates. SNP-variations in the *vanB* gene cluster were compared to increase the comparative resolution. Technical replicates of 2 isolates were sequenced to determine the number of core-SNPs derived from random sequencing errors.

Results. Of the 721 patients screened for VRE carriage, AFLP assigned isolates of 22 patients to the outbreak cluster. According to WGS, all 22 isolates belonged to ST117, but only 21 grouped in a tight phylogenetic cluster and carried *vanB* resistance gene clusters. Sequencing of technical replicates showed that 4-5 core-SNPs were derived by random sequencing errors. The cut-off for the relatedness of epidemiologically linked VRE isolates was established at ≤seven core-SNPs. The discrepant isolate was separated from the index isolate by 61 core-SNPs and the *vanB* gene cluster was absent. In AFLP analysis this discrepant isolate was indistinguishable from the other outbreak isolates, forming a cluster with 92% similarity (cut-off for identical isolates ≥90%). The inclusion of

the discrepant isolate in the outbreak resulted in the screening of 250 patients and quarantining of an entire ward.
Conclusion. AFLP was a rapid and affordable screening tool for characterising hospital VRE outbreaks. For an in-depth understanding of the outbreak, WGS was needed. Compared to AFLP, WGS provided higher resolution typing of VRE isolates with implications for outbreak management.

2. Report on the comparison of NGS versus gold standard methods for metagenomics diagnosis (EMC)

Introduction

Full-genome sequencing of HIV can be challenging due to the high host background and the variability of the viral genome. Sequencing the complete genome will enable clinicians to determine antiviral resistance mutations across the complete genome, but will also enable epidemiologists to track the virus phylogenetically. Various protocols exist to determine HIV sequences. However, these protocols either aim at specific subtypes of HIV or at specific target genes that will not obtain the complete genome.

Abstract

Comparison of full-genome amplicons amplification, metagenomics sequencing and viral metagenomics sequencing of HIV positive plasma samples.
Background. The detection of single nucleotide polymorphisms (SNPs) in HIV for early detection of antiretroviral resistance can be challenging. Current assays are often aimed at specific genes or several genes, therefore missing off-targets mutations that can induce resistance. Additionally, these assays are specific for a particular species or subtype of HIV, making it problematic for HIV-wide SNP detection. We set out to investigate multiple methods that attempt to determine the complete genome of HIV from plasma samples.
Methods. Two plasma samples positive for HIV-1 (2.04x10^5 and 1.26x10^5 c/mL for sample 1 and sample 2, respectively) were selected for a variety of sequencing approaches. We adapted three different approaches to obtain a full-genome sequence of the viruses: (i) PCR amplification of 10 overlapping amplicons and subsequent sequencing of the PCR reaction, (ii) metagenomics sequencing and (iii) viral metagenomics sequencing using virus-specific baits (ViroCap). All libraries were sequenced on an Illumina MiSeq machine. Reads were aligned to the HIV-1 reference genome.
Results. Amplicon amplification and subsequent sequencing resulted in 5.0M and 5.6M reads in total, 67.862 and 88.922 reads aligned to the HIV reference genome, covering 37.4% and 13.1% of the genome. Metagenomic sequencing resulted in 37.9M and 40.5M reads in total, 32.011 and 8.810 reads aligned to the HIV reference genome, covering 45.6% and 22.0% of the genome. Viral metagenomic sequencing resulted in 37.9M and 40.5M reads in total, 2.5M and 1.1M reads aligned to the HIV reference genome, covering 99.9% and 99.4% of the genome (Table 1).

	Amplicon	Metagenomics	VirCapSeq
Total reads			
s1	5.682.130	37.918.335	42.659.318
s2	4.998.217	40.522.959	11.656.686
HIV reads (% total)			
s1	67.862 (11,9)	32.011 (0,1)	2.482.560 (5,8)
*s1 norm.**	*119.431*	*8.442*	*581.950*
s2	88.922 (17,8)	8.810 (0,0)	1.062.470 (9,1)
*s2 norm.**	*177.907*	*2.174*	*911.468*
% genome			
s1	37,4	45,6	99,9
s2	13,1	22,0	99,4

*Normalized to 10M reads

Table 1. Sequencing and coverage results of two plasma samples positive for HIV.

Conclusion. The use of 10 overlapping amplicons did not result in (near) complete genome coverage. Increase in the number of reads would probably not be increasing the total coverage of the genome, for they were very fragmented to certain locations on the genome although most of the reads were specific for HIV. For metagenomics, the coverage increased slightly, but this required a major increase in the total amount of reads to be sequenced because of higher human background sequences. Additionally, pegivirus C was found in sample 2 with 1-2 log more reads as compared to HIV. This, together with the high human background sequences, reduced the total amount of HIV-specific reads dramatically. Viral metagenomics revealed to cover the complete genomes for both samples. The amount of pegivirus C reads also increased in the capture approach, however, the HIV-1 genome was still entirely covered with a large number of reads. We, therefore, conclude that the viral metagenomics approach is highly favoured to obtain full genome sequences from clinical material.

3. Report on the comparison of Biomerieux versus Seqsphere for typing (RKI)

Introduction

Whole-genome sequencing (WGS) is becoming the method of choice for rapid and accurate identification and characterisation of bacteria. However, the challenge remains to translate the data in relevant and easy to interpret epidemiological information. For instance, genotyping information is indispensable for WGS-based phylogenetic analyses (cgMLST) and phenotype prediction (virulence, resistance). To bring such analyses to powerful routine tools are needed that provide efficient workflows and intuitive user interfaces. We compared and validated the genotyping functions of two commercially available software tools (SeqSphere+ and Bionumerics) based on WGS data of almost 300 different *Staphylococcus aureus* strains (Strommenger *et al.*, in prep.; oral presentation at ECCMID 2019, poster presentation at microbe 2019; **Abstract**).

Publication

- Strommenger *et al.*, in prep.
    - Oral presentation at ECCMID 2019, poster presentation at microbe 2019: *NGS-based molecular typing of Staphylococcus haemolyticus - replacing PFGE to increase resolution.*

Fuchs Manuela, Layer Franziska, Bender Jennifer, Thürmer Andrea, Werner Guido, Fuchs Stephan, Birgit Strommenger

- Tentative peer-reviewed manuscript: *Another gaze into the crystal ball - predicting Staphylococcus aureus traits from whole-genome sequencing data*. Scientific Reports.

Abstracts

Background: Here, we describe and validate the use of two commercially available software tools enabling this translation for *Staphylococcus aureus*: SeqSphere+ and the genotyping functionality implemented in the BIONUMERICS software platform.
Methods: A panel of 279 *S. aureus* isolates, representing the diversity of the German *S. aureus* population, was characterized using traditional methods. Antibiotic susceptibility towards 19 antibiotics was determined using broth microdilution. MLST and spa-types were determined using PCR and Sanger sequencing. All isolates were subjected to WGS (Illumina technology) and raw data was assembled using three different assembly algorithms (A5, SPAdes and Velvet). Resulting assemblies were analyzed using SeqSphere+ and the genotyping functionality implemented in BIONUMERICS, respectively. Besides cgMLST/wgMLST, spa-type, MLST and antibiotic resistance were predicted from WGS data and results were compared to those obtained by traditional methods (MIC, PCR, microarray). The Robert Koch-institute (RKI) evaluated the results obtained from SeqSphere+, while the results obtained from BIONUMERICS were evaluated by bioMérieux (bMx). Results were evaluated in a blinded manner, i.e. neither the RKI nor bMx had prior knowledge of the results obtained with the other software.
Results: Regarding spa-type and MLST, high concordance rates (93-99%) between predicted and traditionally obtained results were observed. However, spa-type prediction based on SPAdes assemblies correlated least with the traditionally obtained results, yielding 93% correlation compared to 99% and 97% for Velvet and A5-based predictions, respectively (SeqSphere+). Additionally, concordance rates were lower for spa-type predictions obtained from BIONUMERICS for the SPAdes and Velvet assemblies. For these two assembly algorithms, a new remapping and consensus calling step were performed in Bionumerics after the assembly, which might assure more conservative results but introduces more uncertain bases. The A5 assembly has been performed with the A5-miseq pipeline (not included in either SeqSphere+ or BIONUMERICS), and thus served as precisely the same input in both software packages. Both software packages predict the spa-type equally well based on the A5 assembly.
Two mechanisms play an essential role in antibiotics resistance in *S. aureus*: the acquisition of whole genes conferring resistance, or the accumulation of mutations, leading to a new gene variant coding for increased resistance. For both software packages, almost 90% overall concordance was obtained between traditionally obtained and WGS predicted antibiotic resistance results based on SPAdes and A5 assemblies, whereas only 84% overall concordance was observed for predictions based on Velvet assemblies. Focussing on the prediction of resistance to specific antibiotics, we observe that both software packages perform similarly, although BIONUMERICS performs slightly better in predicting mutation-based resistances. Resistance to fosfomycin proves to be difficult to predict. Although a role of the fosB/fosD gene family has been described in resistance to fosfomycin, no fosB/fosD genes were detected in any of the (four) samples that phenotypically showed resistance to fosfomycin, suggesting another, yet unknown, resistance mechanism.

Conclusion: The use of WGS data in standard bioinformatics pipelines can greatly improve the efficiency and effectiveness of molecular surveillance to predict phenotypic traits. Moreover, these predictions can help to increase our knowledge and understanding of *S. aureus*.