DEVELOPMENT OF A PIPELINE FOR GENOMIC SEQUENCING BASED SURVEILLANCE OF PNEUMOCOCCAL INFECTIONS

Karsten Maruhn[1,2], Andreas Itzek[1], Matthias Imöhl[2] and Mark van der Linden[1]
[1]German National Reference Center for Streptococci, Institute of Medical Microbiology, University Hospital RWTH Aachen
[2]Laboratory Diagnostic Center, University Hospital RWTH Aachen

Karsten Maruhn

German National Reference Center for Streptococci
Institute of Medical Microbiology
University Hospital RWTH Aachen
Pauwelsstrasse 30
52074 Aachen
Germany

+49 241 80 38351
kmaruhn@ukaachen.de

BACKGROUND

The massive parallel sequencing capacity of next generation techniques allows the decoding of the complete DNA sequence of a bacterial organism, a process commonly described as whole genome sequencing (WGS). The huge amount of data, generated by these approaches requires automatized bioinformatics pipelines, for assembly and annotation of the genome, to extract a wide range of organism specific information such as taxonomic classification, clonal lineage, transmission routes, antimicrobial resistance pattern and serotype affiliation.

METHODS

The machine-aided analysis of pneumococcal WGS data included genome assembly, taxonomic classification, identification of virulence factors and *in silico* MLST, using the TORMES program. The assignment of pneumococcal serotypes was addressed by the SeroBA tool.

RESULTS

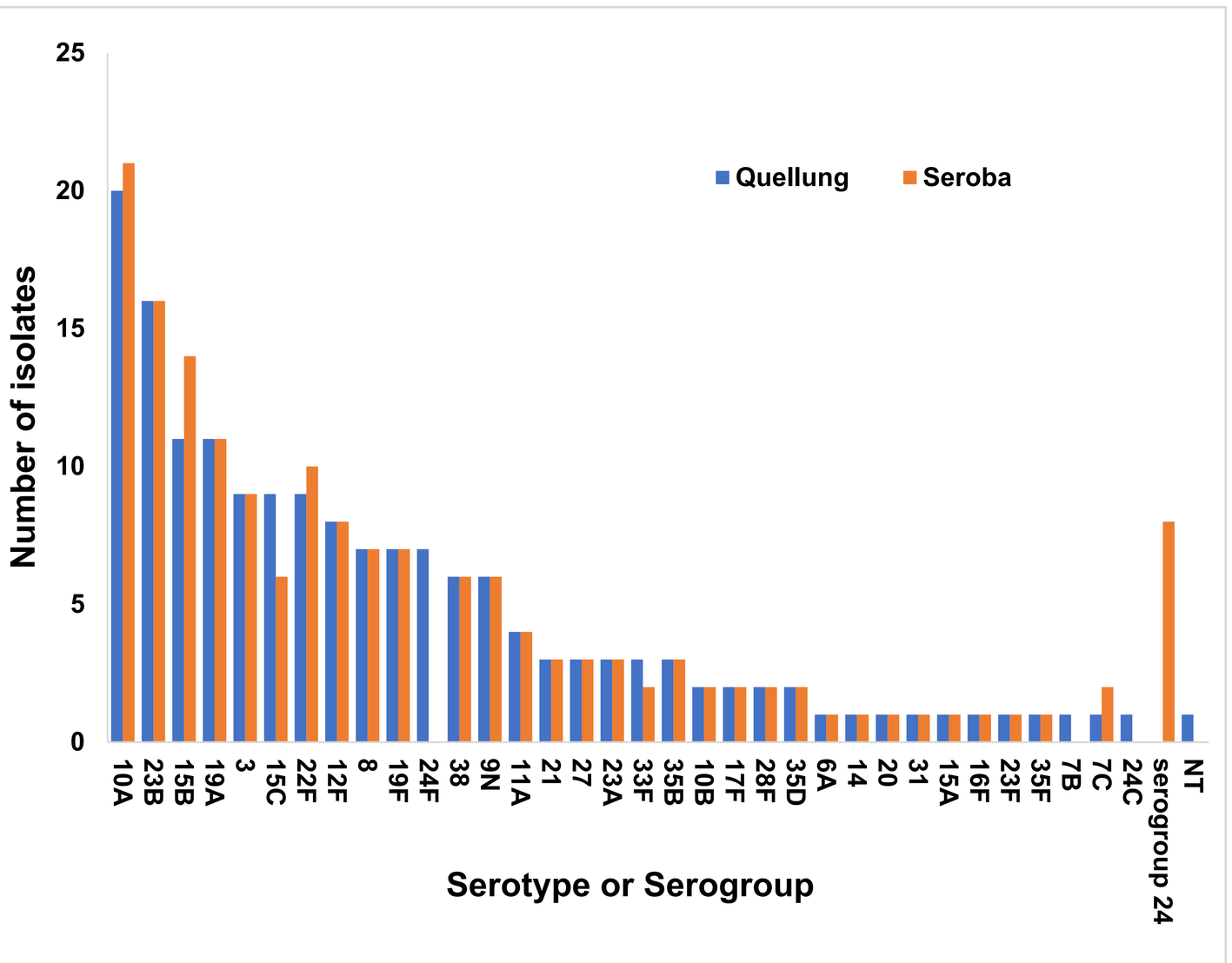
The combination of freely available, community approved solutions with customized scripts allowed convenient semi-automatized processing of WGS data and adaption to specific user needs. Widely scalable parallelism for simultaneous sample analysis was used to save time and optimize computational resources.

First validation runs with WGS data from 165 pneumococcal isolates associated with invasive pneumococcal disease of children in Germany, collected in 2017 and 2018 showed congruent results with taxonomic classification using standard microbiological methods and comparative sequence analyses. *In silico* serotyping was in 96% agreement with the results, obtained from gold standard Quellung reaction (COINCIDENCE OF SEROTYPE DETERMINATION). The WGS based MLST/serotype associations identified by combining the output of TORMES and SeroBA were in agreement with data of our own database and MLST/serotype data available on PubMLST (ASSOCIATION OF SEROTYPE TO MLST).

CONCLUSIONS

The output, generated by the established WGS analysis pipeline was in good agreement with the results obtained using conventional methods, with the benefit of easy handling, quick adaption to changing requirements and reduced laboratory workload. Prospectively, automatized determination of antimicrobial resistance of pneumococcal isolates against antibiotics commonly recommended for the treatment of pneumococcal disease as well as additional serotype prediction algorithms and options for more comprehensive bacterial epidemiological analysis will be implemented.

COINCIDENCE OF SEROTYPE DETERMINATION



ASSOCIATION OF SEROTYPE AND MLST

Serotype	MLST	WGS analysis [n]	GNRCS database [n]	PubMLST database [n]
10A	1551	12	14	59 (1×6B, 18×inconclusive)
	816	5	2	44 (6×10B, 1×19A, 12×inconclusive)
	97	2	7	58 (1×19A, 125×inconclusive)
	473	1	1 (4×6C, 3×6A, 3×15A, 1×6B, 1×15B)	0 (114×14, 2×19A, 1×19F, 1×6B)
23B	1372	6	0	0 (1×22F, 9×inconclusive)
	1373	3	0	82 (2×19A , 1×19F, 4×genetic variant)
	1349	2	10 (2×23A, 1×19A)	23 (1×19A, 1×23A, 8×inconclusive, 2×not determined)
	9867	2	1	6
	1985	1	0 (1×19F)	0 (1×19F, 1×genetic variant)
	11167	1	0 (7×15B, 7×15C, 1×19F)	1 (2×genetic variant)
	NT	1	0	0
15B	1262	4	3 (4×15B, 1×19F)	160 (4×11A, 1×14, 1×15A, 1×6A, 84×inconclusive, 3×not determined)
	8711	3	0 (1×15C)	7 (12×inconclusive, 2×not determined)
	199	2	11 (36×19A, 15×15C, 1×15A)	479 (698×19A, 7×19B, 9×19F, 4×23F 1×15F, 1×23A, 1×7C, 157×inconclusive)
	162	1	0 (53×9V, 3×24F, 2×22F, 2×24C, 1×15C)	33 (214×9V, 3×14, 2×8, 126×inconclusive and other)
	200	1	1	1 (9×14, 6×inconclusive)
19A	994	5	1	91 (1×19C, 17×inconclusive, 1×not determined)
	3546	1	1	7
	667	1	10 (1×6C)	135 (1×6C, 1×14, 1×inconclusive)
	450	1	1	15 (3×inconclusive)
	320	1	12 (5×19F)	686 (101×19F, 2×6B, 1×15BC, 15×inconclusive)
3	180	7	0	1247 (192×inconclusive and other)
	1377	2	3	44 (6×inconclusive)
15C	1262	4	7 (7×15B, 1×19F)	160 (4×11A, 2×19F, 84×inconclusive and other)
	8711	3	1	7 (12×inconclusive)
	199	1	15 (35×19A,11×15B, 4×NT)	479 (698×19A, 9×19F, 4×23F, 7×19B, 157×inconclusive and other)
	1025	1	0	20 (3×19F, 1×3, 1×inconclusive)
22F	433	7	23 (4×35C)	597 (43×23A, 7×19A, 3×42, 3×35C, 2×31, 1×35A, 1×3, 164×inconclusive)
	819	2	4	9 (1×inconclusive, 1×not determined)

Data of the four most common serotypes of the dataset are shown