Establishment and evaluation of a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Pseudomonas aeruginosa*

Hauke Tönnies, Karola Prior, Dag Harmsen, Alexander Mellmann

Supplemental Material (Legends of Supplemental Tables, Supplemental Figures, Supplemental References)

Supplemental Table S1: List of strain, MLST sequence types, and cgMLST allelic profiles used for reproducibility testing of the novel cgMLST scheme. The "R" in the sample ID indicates the second sequencing process starting from a novel colony of the same strain. The cgMLST loci are named in accordance to the locus tags of *P. aeruginosa* reference strain PAO1 (NC_002516). All raw data (fastq) are available under the ENA project accession number PRJEB38241 and the given accession numbers.

Supplemental Table S2: List of core genome genes used for the *P. aeruginosa* cgMLST scheme. Locus tags designations were taken from the *P. aeruginosa* reference strain PAO1 (GenBank acc. no. NC_002516.2).

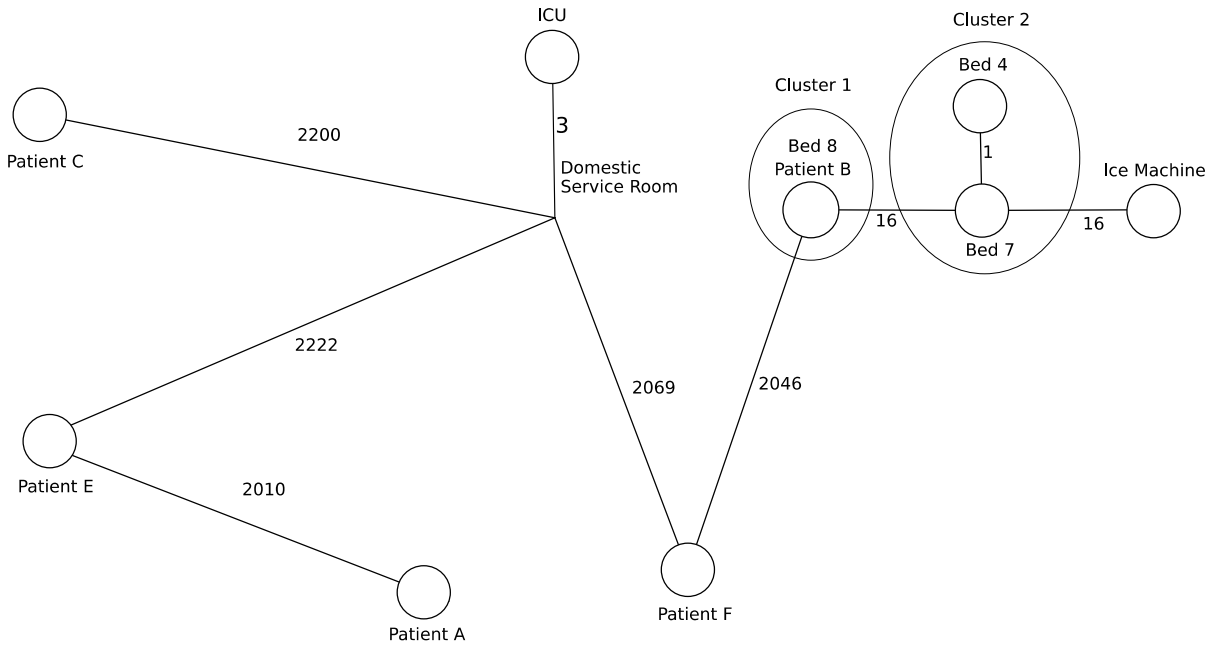
Supplemental Table S3: List of strain, MLST sequence types (ST), and cgMLST allelic profiles of the downloaded fastq files. The cgMLST loci are named in accordance to the locus tags of *P. aeruginosa* reference strain PAO1 (NC_002516.2). All raw data (fastq) were downloaded from the European Nucleotide Archive using the accession number given in the cited publication. n.a., not assigned.

Supplemental Table S4: Each entry of the (symmetrical) matrix consists of the number of differing alleles between the two isolates with the corresponding sample Ids respectively run number shown in the first two columns respectively first two rows. All raw data (fastq) are

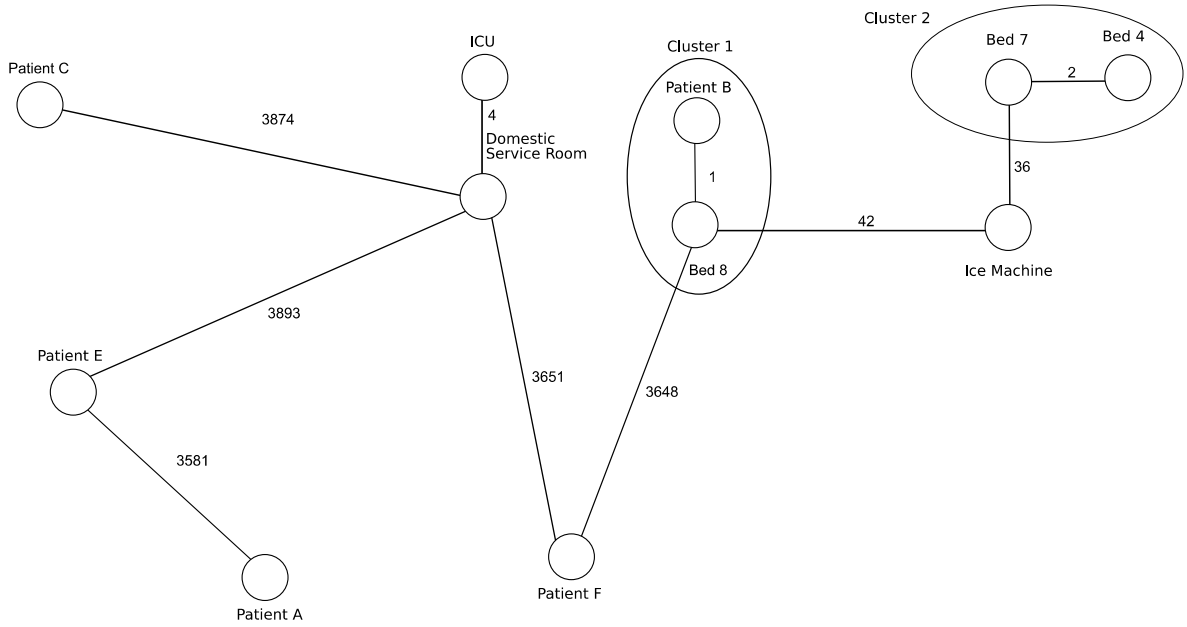
available under the ENA project accession number PRJNA503802 (Magalhães et al. [1]) resp. PRJEB21208 (Parcell et. al. [2]) and the given run numbers.

Supplemental Table S5: The first data sheet lists all 984 successfully downloaded datasets. The second data sheet lists the remaining 922 strains after a fastANI-QC check. Only records with >95% identity to the *P. aeruginosa* type strain (DSM50071T, NZ_CP_012001) were kept. The third data sheet lists the remaining 916 records after MASH screening of potential contamination of these 922 datasets. The fourth data sheet lists the final 490 datasets after removing 426 of the datasets with an assembled coverage < 70x.

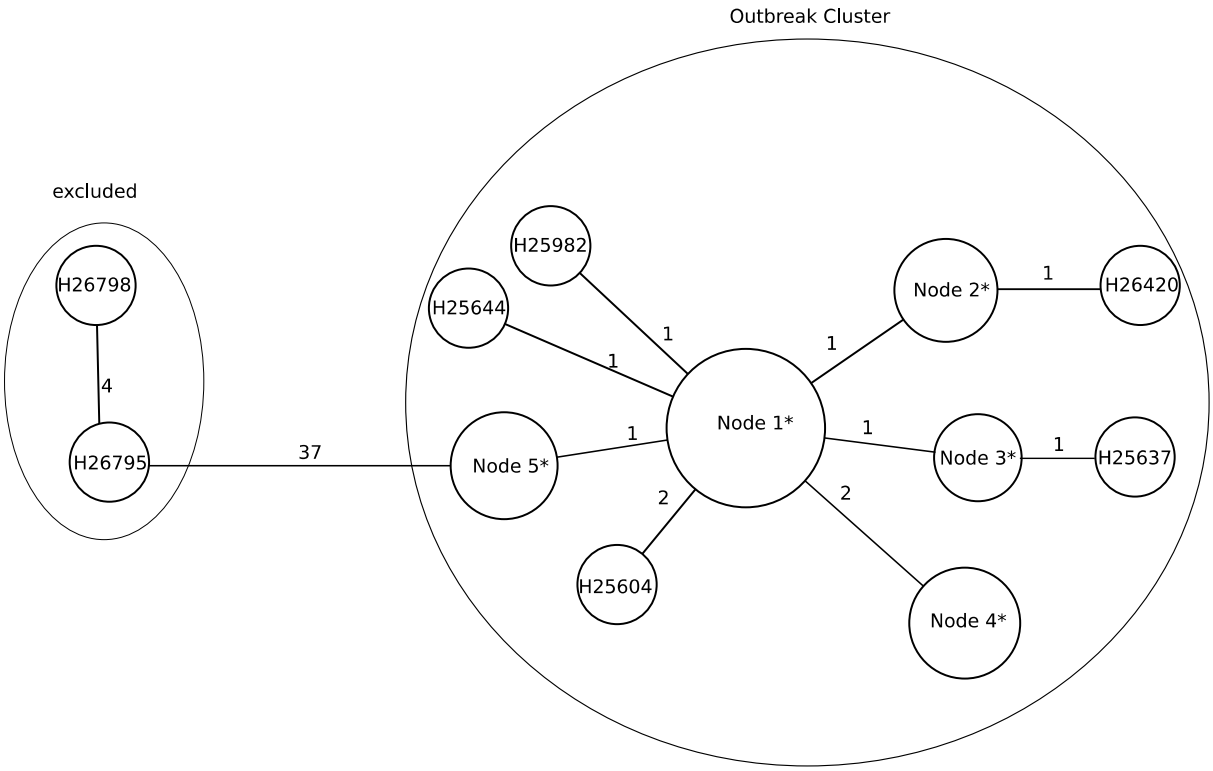
Supplemental Figures



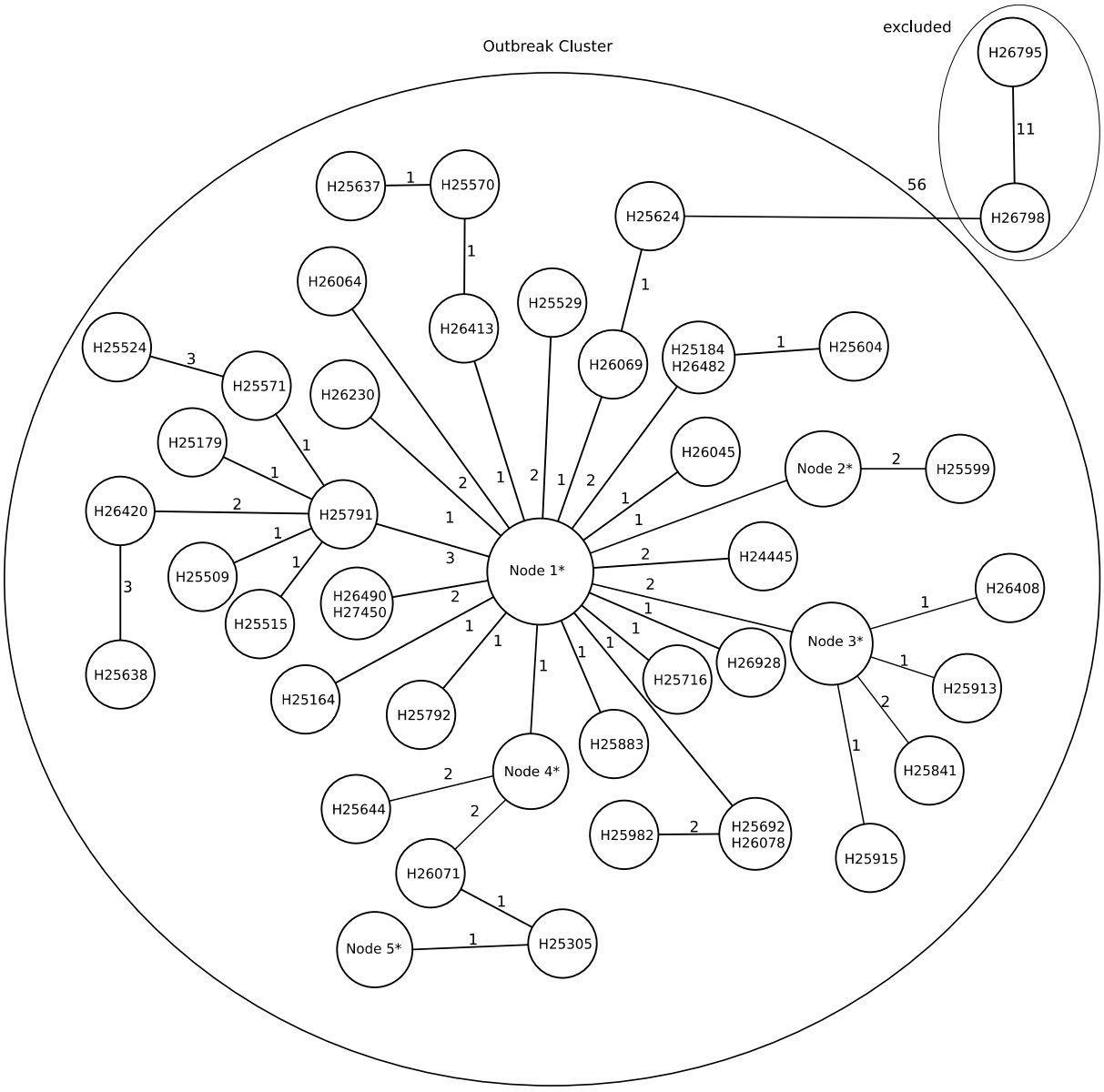
Supplemental Figure S1: Minimum spanning tree based on the allelic profiles of the novel *P. aeruginosa* cgMLST scheme of the genomic sequence data (n= 11 isolates) from Parcell et al. (2) using the scheme of de Sales et al. (3). Each circle represents the genotype based on a unique allelic profile of up to 2,653 cgMLST genes (ignoring missing values in pairwise comparisons), is named with the isolate's label and the number on connecting lines display the number of differing alleles.



Supplemental Figure S2: Minimum spanning tree based on the allelic profiles of the novel *P. aeruginosa* cgMLST scheme of the genomic sequence data (n= 11 isolates) from Parcell et al. (2) using the scheme of Stanton et al. (4). Each circle represents the genotype based on a unique allelic profile of up to 4,440 cgMLST genes (ignoring missing values in pairwise comparisons), is named with the isolate's label and the number on connecting lines display the number of differing alleles.

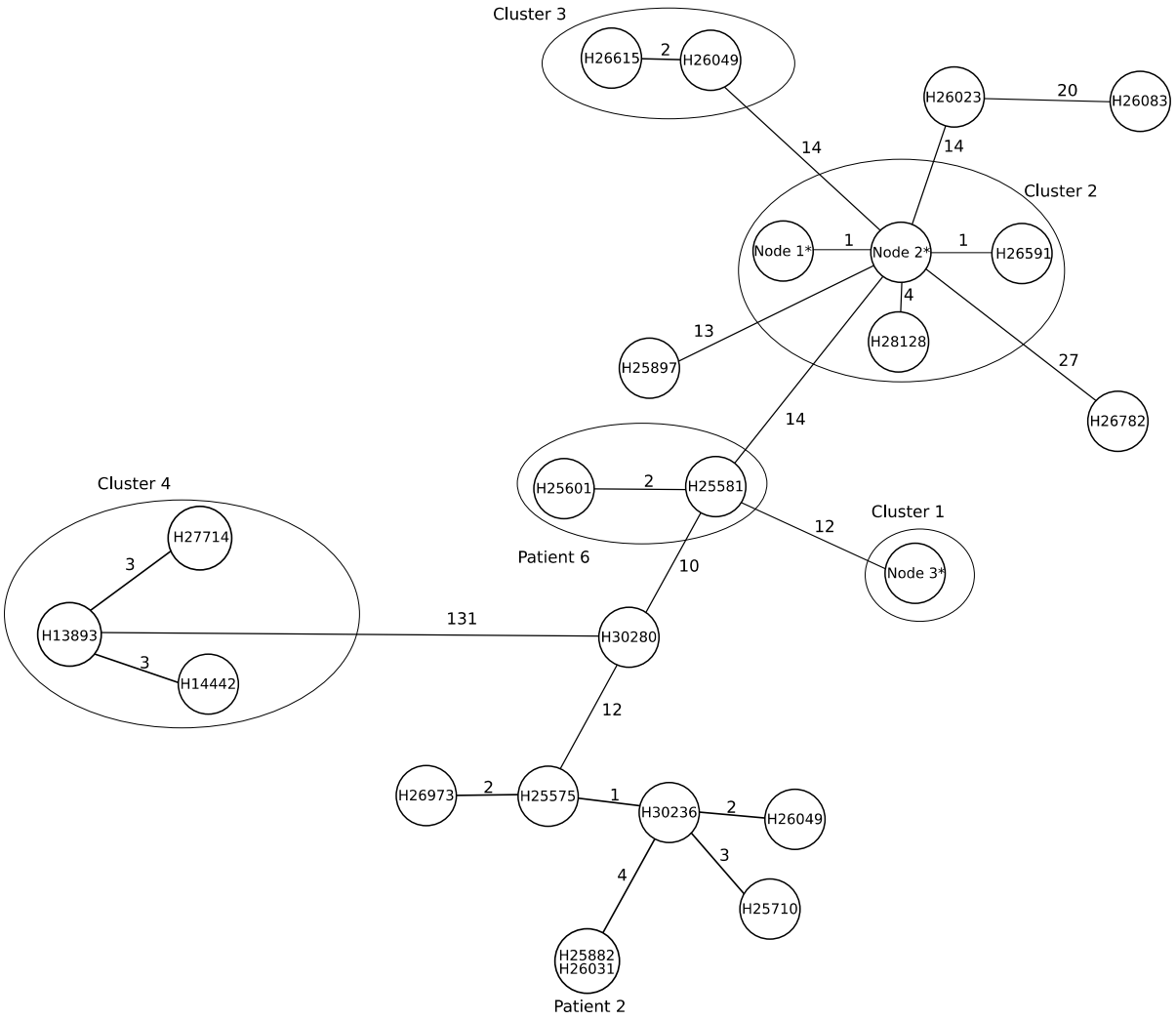


Supplemental Figure S3: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=67) with the MLST ST1076 gathered from Magalhães et al. (1) using the scheme of de Sales et al. (3). Each circle represents the genotype based on a unique allelic profile of up to 2,653 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates. If more than two isolates belong to the same node, the node is marked with an asterisk (*) comprising the following isolates: Node 1: H24445, H25184, H25305, H25305, H25328, H25469, H25471, H25473, H25525, H25529, H25599, H25624, H25688, H25692, H25706,H25716, H25776, H25954, H26060, H26069, H26071, H26076, H26078, H26166, H26188, H26482, H26490, H26927, H26928, H26929, H26932, H26934, H26935, H27450; Node 2: H25179, H25509, H25515, H25524, H25571, H25638, H25791; Node 3: H25570, H25792, H26413; Node 4: H25689, H25727, H25841, H25905, H25913, H25915, H26073, H26408, H26410; Node 5: H25162, H25163, H25164, H25784, H25883, H26045, H26064, H26230

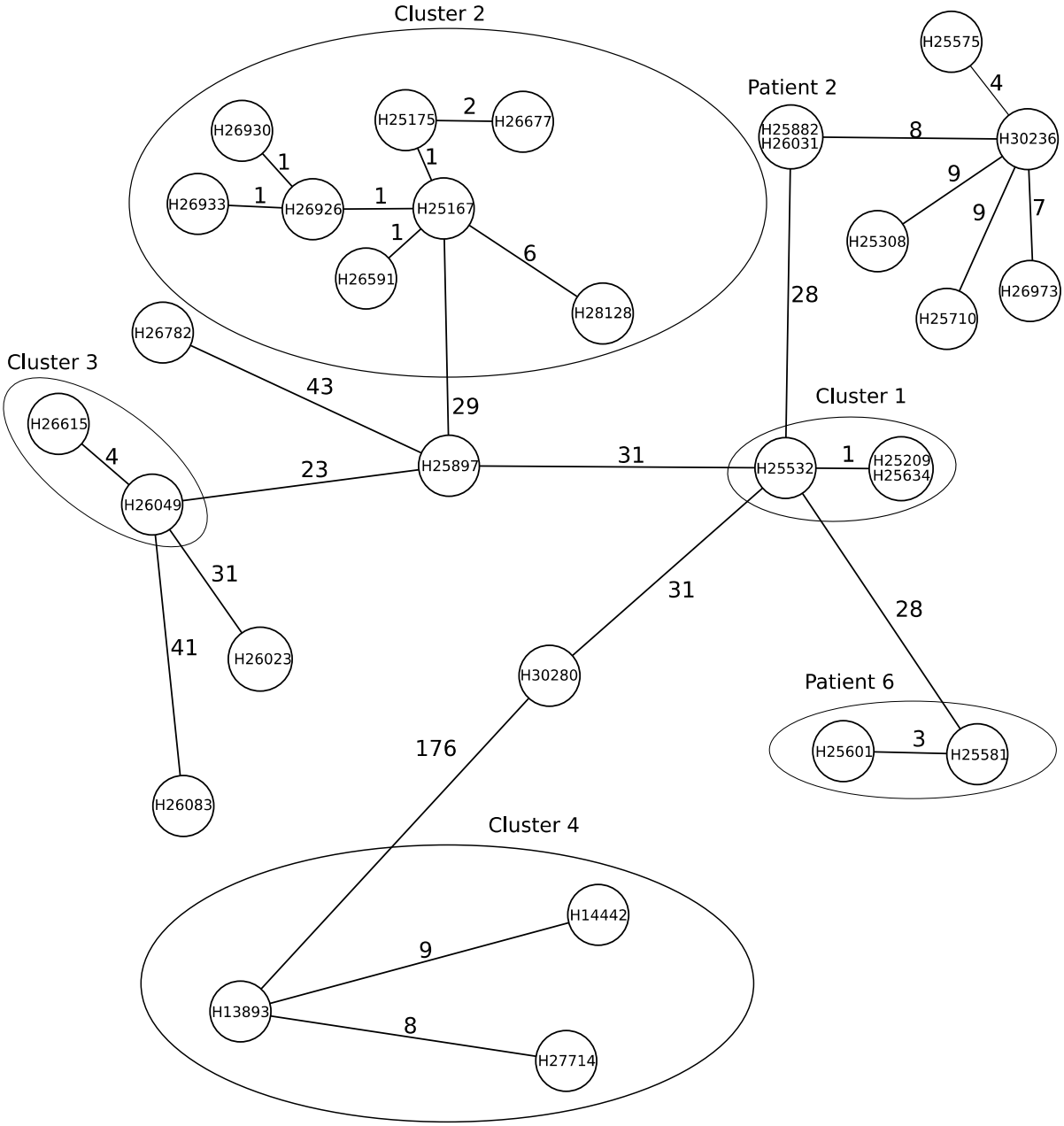


Supplemental Figure S4: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=67) with the MLST ST1076 gathered from Magalhães et al. (1) using the scheme of Stanton et al. (4). Each circle represents the genotype based on a unique allelic profile of up to 4,440 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates. If more than two isolates belong to the same node, the node is marked with an asterisk (*) comprising the following isolates: Node 1: H25162, H25163, H25776, H25784, H25954, H26060, H26076, H26927, H26929, H26932, H26934, H26935;

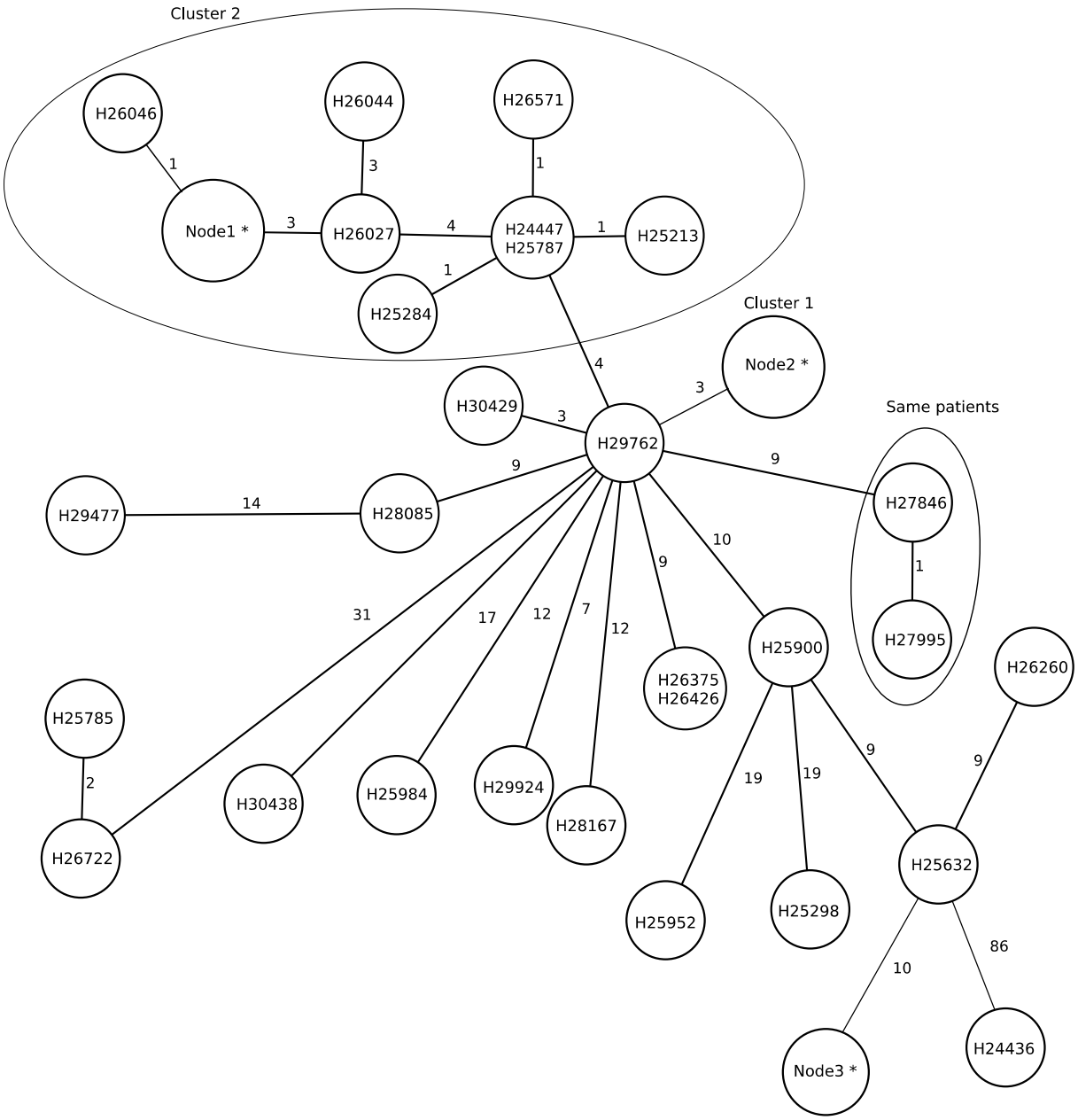
Node 2: H25473, H25688, H25706; Node 3: H25689, H25727, H25905, H26073, H26410; Node 4: H25328, H25469, H25525; Node 5: H25471, H26166, H26188



Supplemental Figure S5: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=31) with the MLST ST253 gathered from Magalhães et al. (1) using the scheme of de Sales et al. (2). Each circle represents the genotype based on a unique allelic profile of up to 2,653 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates. If more than two isolates belong to the same node, the node is marked with an asterisk (*) comprising the following isolates: Node 1: H26929, H26930, H26933; Node 2: H25167, H25175, H26677; Node 3: H25209, H25532, H25634

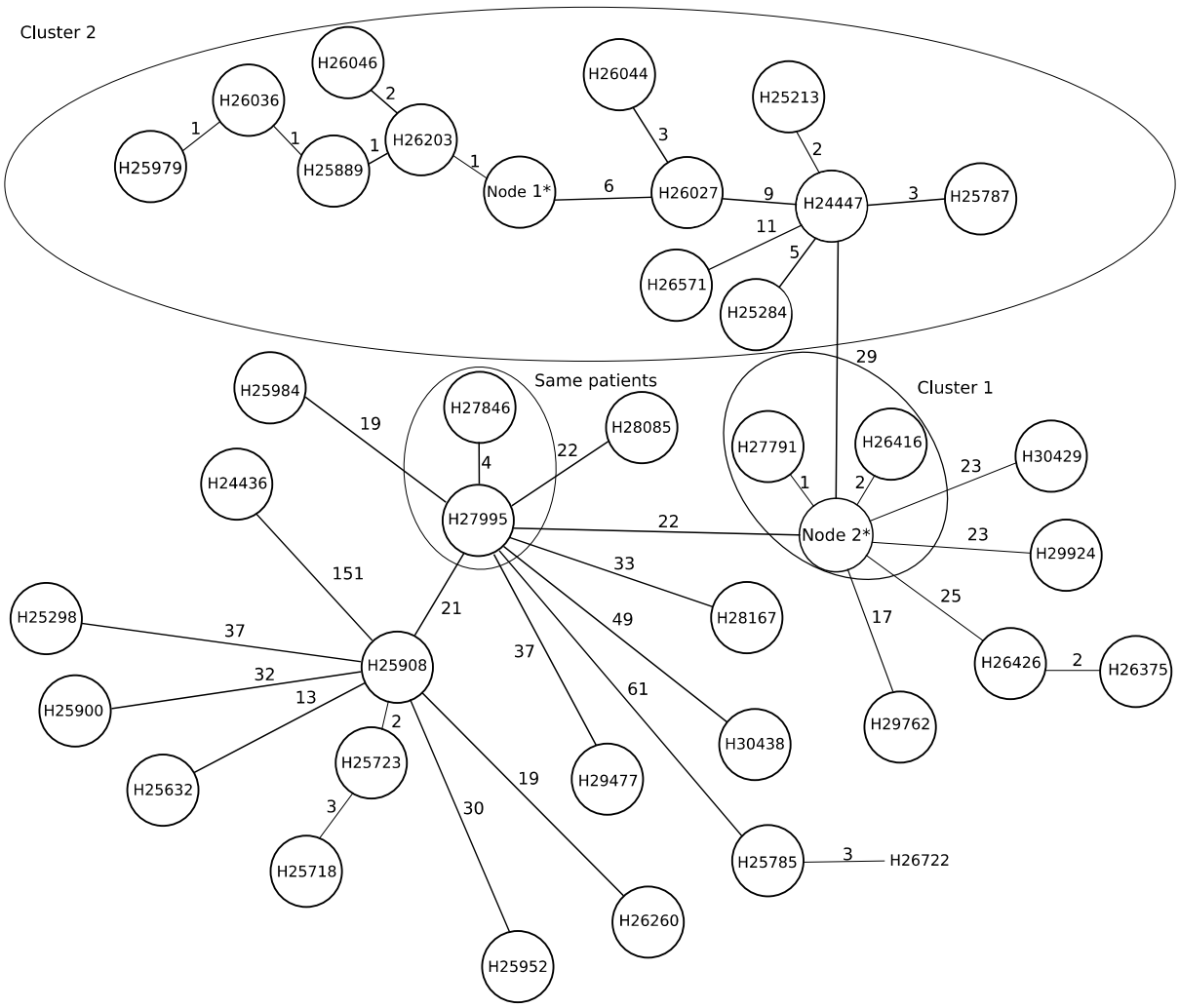


Supplemental Figure S6: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=31) with the MLST ST253 gathered from Magalhães et al. (1) using the scheme of Stanton et al. (4). Each circle represents the genotype based on a unique allelic profile of up to 4,440 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates.



Supplemental Figure S7: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=45) with the MLST ST17 gathered from Magalhães et al. (1) using the scheme of de Sales et al. (2). Each circle represents the genotype based on a unique allelic profile of up to 2,653 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates. If more than two isolates belong to the same node, the node is marked with an asterisk (*) comprising the following isolates: Node 1: H25889, H25979,

H26036, H26084, H26086, H26202, H26203; Node 2: H25200, H25508, H25961, H26247, H26416, H26524, H27791; Node 3: H25718, H25723, H25908



Supplemental Figure S8: Minimum spanning tree based on the allelic profiles of the genome data of all isolates (n=45) with the MLST ST17 gathered from Magalhães et al. (1) using the scheme of Stanton et al. (4). Each circle represents the genotype based on a unique allelic profile of up to 4,440 cgMLST genes (ignoring missing values in pairwise comparisons) and the number on connecting lines display the number of differing alleles. The circles are named with the isolates. If more than two isolates belong to the same node, the node is marked with an asterisk (*) comprising the following isolates: Node 1: H26084, H26086, H26202; Node 2: H25200, H25508, H25961, H26247, H26524

Supplemental References

1. Magalhães B, Valot B, Abdelbary MMH, Prod'hom G, Greub G, Senn L, Blanc DS. 2020. Combining standard molecular typing and whole genome sequencing to investigate *Pseudomonas aeruginosa* epidemiology in intensive care units. Front Public Health 8:3. doi:10.3389/fpubh.2020.00003.
2. Parcell BJ, Oravcova K, Pinheiro M, Holden MTG, Phillips G, Turton JF, Gillespie SH. 2018. *Pseudomonas aeruginosa* intensive care unit outbreak: winnowing of transmissions with molecular and genomic typing. J Hosp Infect 98:282–288. doi:10.1016/j.jhin.2017.12.005.
3. De Sales RO, Migliorini LB, Puga R, Kocsis B, Severino P. 2020. A core genome multilocus sequence typing scheme for *Pseudomonas aeruginosa*. Front. Microbiol. 11:49. doi:10.3389/fmicb.2020.01049.
4. Stanton RA, McAllister G, Daniels JB, Breaker E, Vlachos N, Gable P, Moulton-Meissner H, Halpin AL. 2020. Development and application of a core genome multilocus sequence typing scheme for the healthcare-associated pathogen *Pseudomonas aeruginosa*. J Clin Microbiol. doi:10.1128/JCM.00214-20.