A Population-level Strain Genotyping Method to Study Pathogen Strain Dynamics in Human Infections.

Sarah J Morgan[a1], Samantha L Durfey[a1], Sumedha Ravishankar[a], Peter Jorth[b], Wendy Ni[a], Duncan Skerrett[a], Moira L Aitken[c], Edward F Mckone[d], Stephen J Salipante[c], Matthew C Radey[a] & Pradeep K Singh[a,b#]

[a] Department of Microbiology, University of Washington School of Medicine, Seattle WA, USA.
[b]Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA
[c] Department of Medicine, University of Washington School of Medicine, Seattle WA, USA.
[d] St. Vincent's University Hospital, Dublin Ireland
[e] Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle WA, USA.

Running title: Population-level Strain Genotyping

Keywords: *Pseudomonas aeruginosa*, *Staphylococcus aureus*, Multi-locus Sequence typing, Cystic Fibrosis, population, diversity, chronic infection

[#] Address correspondence to: Pradeep Singh singhpr@uw.edu

[1] Sarah Morgan and Samantha Durfey contributed equally to this work.

Abstract

A hallmark of chronic bacterial infections is the long-term persistence of one or more pathogen species at the compromised site. Repeated detection of the same bacterial species can suggest that a single strain or lineage is continually present. However, infection with multiple strains of a given species, strain acquisition and loss, and changes in strain relative abundance can occur. Detecting strain-level changes and their effects on disease is challenging as most methods require labor intensive isolate-by-isolate analyses, thus, only a few cells from large infecting populations can be examined. Here we present a population-level method for enumerating and measuring the relative abundance of strains called "PopMLST". The method exploits PCR amplification of strain-identifying polymorphic loci, next-generation sequencing to measure allelic variants, and informatic methods to determine whether variants arise from sequencing errors or low abundance strains. These features enable PopMLST to simultaneously interrogate hundreds of bacterial cells that are either cultured *en masse* from patient samples, or are present in DNA directly extracted from clinical specimens without *ex vivo* culture. This method could be used to detect epidemic or super-infecting strains, facilitate understanding of strain dynamics during chronic infections, and enable studies that link strain changes to clinical outcomes.

Introduction

Serial culturing of chronic infection sites often repeatedly yields the same pathogen species. For instance, chronic wounds can consistently grow *Staphylococcus* and *Pseudomonas* species (1), subjects with urinary tract anomalies can be persistently infected by *Escherichia coli* (2), and chronically-infected sinuses can recurrently yield the same anaerobes (1-3). The chronic infections that afflict people with cystic fibrosis (CF) are a prime example, as the same pathogen species are frequently cultured from patients' lung secretions for long periods. Some species, like *Pseudomonas aeruginosa* (Pa) and *Staphylococcus aureus* (Sa) can be highly abundant in the lungs of individual patients for decades or even life-long (4-9).

Repeated detection of the same bacterial species over time can imply that a single strain or lineage is continually present. However, even though most stain-level genotyping studies examine very few isolates from each infection, studies on chronic wound, urinary tract, ear, gastrointestinal, and lung infections suggest more complexity. For example, stain-level genotyping methods have shown that close to a third people with CF and Sa lung infections simultaneously harbor more than one Sa strain (4, 5, 10, 11). Likewise, up to 40% of people with CF and Pa lung infections are simultaneously infected by more two or more Pa strains (12-14), although other work has suggested lower frequencies (7, 15-19). In addition, strain relative abundance can change over time, and strains can be gained or lost in individual patients (13, 18, 20). Notorious examples are Pa epidemic strains that can infect and eventually become dominant in already-colonized patients, and markedly worsen disease (21-23).

Identifying infecting strains is important for several reasons. First, strains of the same species can differ markedly in traits like the capacity for injury, transmissibility, and resistance to antibiotics (19, 22-27). Thus, the presence of multiple strains or changes in strain relative abundance could have clinical consequences. Second, strain abundance changes could provide information about the status of host defenses, treatment efficacy, or pathogen functioning. For example, strains may recede when host defenses or treatments to which they are susceptible intensify, or when deleterious mutations arise. Likewise, new strain acquisition could indicate that host conditions have become more permissive, and analysis of succeeding strains could increase understanding of bacterial functions important *in vivo*. Third, sensitive methods for strain detect could reveal outbreaks and lapses in infection control procedures. Finally, early detection of new strains could spur eradication attempts, which may be more successful soon after strains are acquired (28-32).

Established methods for strain-level identification such as pulse-field gel electrophoresis (PFGE), multi-locus sequence typing (MLST), whole genome sequencing (WGS), and others must generally be performed on one cultured isolate at a time (16, 33-36). Because pathogen populations can be extremely large and colonies from different strains may look identical (37-41), analyzing a few colonies per sample could miss multi-strain infections and strain acquisition and loss events. Newer methods using amplification of species-specific variable regions (42, 43) are not easily adaptable to multiple pathogens, and shotgun sequencing of clinical samples (41, 44) can be limited if non-target DNA (e.g. host or other bacterial DNA) is abundant. To address these limitations, we developed PopMLST ("population MLST"), a method to enumerate and measure the relative abundance of strains present in pools of hundreds of cultured isolates, or in DNA directly extracted from clinical samples.

Results

Overview. In conventional MLST, bacterial colonies are isolated in pure culture, Sanger sequencing is used to identify allelic variation in MLST loci within conserved housekeeping genes (7 loci in the case of Sa and Pa), and loci allele types are determined by comparison to a database (45-47). Because a single clone is analyzed, the loci are known to be linked, in that they originate from the same bacterial isolate. Thus, loci allele identities can be combined to define the MLST type of a pure culture isolate.

In contrast, the goal of PopMLST is to enumerate the pathogen strains and measure strain relative abundance in samples that contain multiple strains and a vast excess of non-target (e.g. human) DNA. To enable this, PopMLST uses PCR to amplify MLST loci from complex samples, and next-generation sequencing to measure allele relative abundance. The PCR primers act as probes to find conserved sequences flanking MLST loci (even when the targeted species is rare), and as vectors to amplify the strain-discriminating MLST loci. Amplicons are Illumina sequenced, the bioinformatic tools are used to distinguish rare variants from errors, bin "like" sequences, and measure their relative abundance (Figure 1).

A drawback to this approach is that PCR amplification and Illumina sequencing from complex mixtures is more error-prone than Sanger sequencing of individual clones, and errors could be confused for low-abundance variant strains. We addressed this problem in several ways. (1) We used high fidelity polymerases and as few PCR cycles as possible to reduce error and PCR chimeras. (2) We adapted the DADA2 analysis pipeline (48) designed for 16S rRNA amplicon sequencing that uses statistical methods to distinguish sequencing errors from low abundance

variants (Figure 1B)) (43, 48). (3) We developed bioinformatic methods to adaptively trim the lower-quality ends of the second read generated by Illumina sequencing to facilitate accurate read merging (Figure 1B and methods). (4) We amplified each MLST locus in triplicate and pooled the data to reduce random, preferential amplification of templates (i.e. "jackpot" amplifications) (49). (5) We omitted a GC-repeat rich Pa MLST loci (*aro*) that was challenging to sequence with Illumina chemistries (Figure S1 A-C) (50, 51), as we found that 91% of 3,379 MLST types in the MLST database (47) could be identified without *aro* (Figure S1D). Together these approaches mitigate, but do not fully eliminate the effects PCR and sequencing errors.

Data interpretation. While the PCR and Illumina sequencing used in PopMLST enable analysis of complex mixtures containing multiple stains and excess non-target DNA, information from MLST loci are unlinked, as many (up to hundreds) of isolates are analyzed *en mass* and sequence reads reporting alleles from each locus are derived from separate PCR reactions. This issue does not generally limit PopMLST's ability to enumerate and measure strain relative abundance, which can be ascertained by examining the loci with the highest number of alleles represented. This approach is effective because even though strains sometimes share MLST alleles (and PopMLST will report the sum of the shared loci's relative abundance in these cases), the large number of alleles for each locus (e.g. Sa MLST loci have 484-892 distinct alleles, and Pa MLST loci have 137-278 distinct alleles) make it unlikely that strains would have identical alleles at enough MLST loci to prevent strain enumeration.

The MLST types of strains within mixtures can also often be determined from PopMLST data. When a limited number of strains co-exist, inference can determine which MLST alleles

originate from the same strain, as linked alleles will be detected at a similar relative abundance. For example, if popMLST finds that each loci contains 3 alleles at a relative abundance of 70% : 25%: 5%, it is likely that the alleles identified at 70% relative abundance belong to one strain, alleles at 25% come from a second strain, and alleles at 5% come from a third strain. When many strains are present, if strains co-exist at similar relative abundances, or if strains happen to share several alleles, inference can fail. If knowledge of the specific MLST types is important, conventional MLST can be performed on few cultured colonies to determine which alleles are linked to one another to guide analysis of population-level data generated by PopMLST.

PopMLST identifies single strains after *in vivo* diversification. As an initial test of the method, we performed PopMLST on pure cultures of Sa and Pa and found that >99% of reads correctly reported a single MLST type in each of 21 independent experiments (Table 1).

In CF and other chronic infections, strains genetically diversify during infection (10, 19, 25-27, 52, 53), and within-strain genetic diversity could be mistaken for strain differences. Thus, we tested PopMLST on pools of 90-96 clonally-related Pa isolates collected from different lung regions, from three CF patients undergoing lung transplantation. Whole genome sequencing showed that isolates from each subject were clonally-related to each other, but had genetically diversified via *in vivo* evolution (19). Indeed, two of the three collections exhibited hypermutator phenotypes due to mutations in either *mutL* and *mutS* mismatch repair genes (19). Core genomes of 96 isolates from the subject that was not a hypermutator contained a total 328 SNP differences, and the 96 isolates from subjects with hypermutator lineages contained 3169 and 1653 SNP differences (19).

Despite this extensive diversity, PopMLST correctly identified each of the populations as containing a single MLST type (< 0.01% of reads erroneously reported a second MLST allele) (Figure 2). These data suggest that the measures used to mitigate PCR amplification and sequencing errors are effective for pure-culture isolates and diversified clonally-related populations.

PopMLST accurately measures pathogen strains in experimental mixtures. A key assumption of our approach is that the relative abundance of MLST loci present in samples is maintained through DNA extraction, amplification, sequencing, and enumeration steps (Figure 1). We therefore began testing PopMLST's ability to detect multiple strains using defined mixtures of purified DNA from different strains. PopMLST identified the expected ratios (within 2-fold) of mixtures containing two Sa or Pa strains over a wide relative abundance range (Figure 3). Replicate experiments using different sequencing runs and different MLST types produced similar results (Figure 3A-D and S2-S3). Linear regression of data from the experimental mixtures indicated close agreement between observed and expected findings ($R^2 = 0.9916$ for Sa and $R^2 = 0.9901$ for Pa) with slopes approximating 1 (Sa: 1.017 [95% CI: 0.9872-1.047]; Pa: 0.9806 [95% CI: 0.9454-1.016]). PopMLST was also able to measure relative abundances of three and four strain mixtures (Figure 3E and F).

Despite use of triplicate PCR reactions a single locus type was occasionally detected at higher than expected abundance (Figures 3 and S2-S3). These finding are likely due to PCR bias (indicated by ‡ in Figures 3 and S2-S3) or jackpot amplifications (indicated by # in Figures S3),

which also occurs in 16S rRNA gene measurements (54) which also uses amplicon sequencing. However, because PopMLST integrates data from 6 or 7 independently-amplified loci (unlike 16S sequencing which relies on a single locus), loci that appear to be outliers can be interpreted in context of others to estimate strain relative abundance. Thus, PCR bias did not affect the number of strains detected or markedly affect strain relative abundance in the known strain mixtures (Figure 3, Figure S2-S3).

PopMLST has a low frequency of false positive strain calls. Error inherent to PCR and Ilumina sequencing could cause PopMLST to artifactually report strains that are not present. We reexamined control experiments containing between 1-4 strains of known composition (n=38 for Pa and n=41 for Sa) to examine the effect of using different abundance thresholds to make strain presence and absence calls. As shown in Table 2, using the criterion that single variant locus be present at ≥1% relative abundance falsely registered the presence of a new strain in 9/49 (18%) of control experiments with Pa, and 7/41 (17%) of control experiments with Sa. The criterion that two or more loci be present at ≥1%, or raising the relative abundance threshold for a single allele to ≥4% produced accurate calls in all 49 Pa, and all 41 Sa experiments. We conclude that detection of a single variant loci at greater that 4% relative abundance or two variant loci at >1% relative abundance can be used as a reliable marker of strain presence.

PopMLST can detect specific MLST types with high sensitivity. In certain settings, clinicians and researchers need to detect specific strains with a known MLST type. Examples include superinfections with virulent Pa epidemic strains in people with CF already colonized by Pa, or infection control surveillance during outbreaks. Theoretically, known MLST types should be

detectable with much higher sensitivity than unknown types, as it is extremely unlikely that the chance occurrence of errors would report the presence of the specific MLST loci of interest.

To test this, we measured PopMLST's sensitivity to detect targeted low abundance MLST alleles in complex mixtures. As shown in Table 3 and Table 4, targeted low abundance alleles were detected in all experiments when present at 2% relative abundance or greater, and in almost all experiments when present at 1% and 0.1% relative abundance. These findings suggest that PopMLST could be used for early detection of known strains with high transmissibility or virulence, or to investigate efficacy of infection control measures.

PopMLST works in the presence of excess human or non-target bacterial DNA. Clinical samples can contain vast amounts of human and non-target bacterial DNA. For example, despite high pathogen density (Pa can reach 10^8-10^9 CFU/ml in CF sputum), 95-99% of CF sputum DNA is human (55), and DNA from other pathogens or oral bacteria can also be highly abundant.

We investigated the effects of contaminating DNA on PopMLST two ways. First, we preformed PCR on human and non-target bacterial DNA (including closely-related species) using Sa and Pa PopMLST primers. Amplicon yields and the number of reads mapped to Sa and Pa MLST loci in these experiments were similar to no-template controls (Figure 4A-D). Second, we tested the ability of PopMLST to detect strains in the presence of 95% human DNA, and found that the vast excess of human DNA did not compromise detection, even when strains were present at as low as 1% relative abundance (Figure 4E-F).

PopMLST measures strain abundance in clinical samples. Encouraging results with experimental strain mixtures led us to perform proof-of-principle tests of PopMLST on clinical samples. In the first test, we cultured sputum from seven Sa-infected CF subjects, and performed PopMLST on DNA prepared from ~100 colonies that grew from each sample (scraped *en mass* from culture plates). PopMLST reported that three of seven samples contained two Sa MLST types (Figure 5A). We tested these results by Sanger sequencing a distinguishing MLST locus in 20-30 individual colonies from each sample, and found MLST types at a similar relative abundance as determined by PopMLST ($R^2 = 0.9247$; slope = 0.9296 [95% CI: 0.5612-1.298]) (Figure 5A-B).

Second, we tested PopMLST directly on CF sputum (without culturing isolates) by mixing two sputum samples that were known to harbor different Pa strains. Each sputum sample contained a single strain by popMLST, and both strains were detected in the mixtures of the two sputum samples (Figure 5C).

Finally, we analyzed sputum samples from 5 subjects that were all known to harbor 2 Sa strains each. We performed PopMLST on DNA prepared directly from sputum and from ~100 Sa isolates from each sample scraped *en mass* from culture plates, and compared results to Sanger sequencing select MLST loci from individual cultured isolates (as above). As shown in Figure 5D, PopMLST detected the same MLST types in sputum DNA and culture scrapes from all 5 subjects. In two subjects, strain relative abundance differed significantly in sputum as compared to the culture scrapes (Subject 4 and 5 $p < 0.001$ by multiple t-test). This finding could be due to

differing sensitivity of strains to host defenses (i.e. more of one strain's DNA originates from dead cells), or differential growth capacity of strains in *ex vivo* culture conditions.

Discussion

Infecting bacteria can exhibit heritable diversity at the species, strain, and intra-strain level as diversity can evolve within lineages during infection. While recent findings and new methods have accelerated work on species diversity (56-64) and the diversification of a single strain (10, 19, 25-27, 52, 53), studies of strain-level diversity have lagged. A major factor is that established methods to detect strains must generally be performed on one cultured isolate at a time (16, 33-36). Thus, non-dominant strains are difficult to detect.

PopMLST addresses this limitation, as it can estimate the relative abundance of strains in pools of tens to hundreds of isolates that have been cultured from infected sites, and can be used on DNA extracted directly from clinical specimens without prior culture. Other advantages include inherent technical approaches to minimize PCR and sequencing error, its robustness when human or other bacterial DNA is in vast excess (Figure 4), its ability to detect targeted strains at very low relative abundance (Table 3 and 4) and strains with *in vivo* growth defects, when the method is used on DNA prepared directly from clinical specimens. Furthermore, the method is accurate even when intra-strain genetic diversity has evolved (Figure 2), likely because MLST loci are within conserved "housekeeping" genes that are less variable than elements involved in Spa typing (6), RAPID or PFGE (16, 33). Finally, MLST databases are in widespread use and exist for over 100 bacterial species (47), so PopMLST can be easily be adapted for use for many organisms and results are comparable between laboratories.

PopMLST also has several limitations. First, PopMLST cannot distinguish between unrelated strains having the same MLST sequence type (11, 35), although co-infection with such strains is

unlikely for pathogens for which many MLST types have been identified (there are ~3,500 Pa and ~ 5,500 Sa described MLST types to date (47)). Second, while PCR and Illumina sequencing enable the method to be used on complex mixtures containing multiple strains and abundant non-target DNA, these techniques are subject to errors and biases. We reduce, but cannot entirely eliminate, the effect of these problems using replicate PCR, adaptive trimming, and statistical methods which enable strains to be detected at ~1-3% relative abundance. Third, PopMLST does not identify which MLST loci are linked in individual isolates. While this limitation does not compromise strain enumeration and relative abundance measurements under most circumstances, it can sometimes complicate identification of strain types present.

We expect that the use of PopMLST could increase understanding of strain dynamics during infections. Patients with chronic bacterial infections frequently experience highly variable disease manifestations, treatment responses, and rates of progression. Such variation can be seen between individual patients infected with the same pathogen(s), or in individual subjects at different times even when no change in infecting pathogen species occurs. The acquisition of different strains of a given species or changes in strain relative abundance could account for some of this variation (65). PopMLST will enable tests of new hypotheses exploring the effects of strain-level diversity on treatment resistance and disease manifestations. The methods could also be useful for detecting infection outbreaks in already colonized patients, for testing the adequacy of infection control procedures.

Methods

Patient samples. Sputum samples were collected in accordance with University of Washington Institutional Review Board (protocols numbers 06-4469 and STUDY00011983), and St. Vincents Hospital, Dublin IE (RS20-048). Patients provided written informed consent prior to collection of samples. Sa was isolated after sputolysin diluted sputum was cultured on Mannitol Salts Agar (Difco). Populations were scraped from plates containing >100 colonies by flooding the plate with 2 ml of LB and using a L-spreader to resuspend the bacteria. Pa was isolated after sputolysin diluted sputum was cultured on MacConkey (Difco). All cultures were stored at -80^{o}C in 15% glycerol prior to analysis.

DNA isolation. DNA extraction of Sa samples was performed using the DNeasy Power Soil Kit (Qiagen) with the following modifications: samples were incubated with 2.9 mg lysozyme and 0.14mg lysostaphin prior to lysis with 0.1 mm beads using bead beater (Mini-Beadbeater-16; Biospec) (41). Pa DNA was isolated from 100 ul of resuspended culture using the DNeasy Blood and Tissue kit (Qiagen) using the protocol for gram negative bacteria. DNA was isolated from 100 ul of fresh sputum using the Microbiome Kit (Qiagen).

Control mixtures. Control strains (Table S1 and Table S2) were streaked from freezer stocks and an isolate was grown overnight prior to DNA isolation. Cultured HELA or HEK293 cells were pelleted and DNA was isolated as above. Isolated DNA was quantified by Qubit, and mixed at ratios described in the figures. MLST types of control strains were based on the MLST database and confirmed by MLST typing of single isolates if necessary (Table S1 and Table S2).

PAO1-lacZ:PA14 mixtures were pre-mixed at designated ratios and plated on LB+xGal to confirm the ratio. Growth from the plate was scraped and subjected to DNA isolation as above.

Amplification and sequencing with PopMLST. 5 ng/ul DNA from cultured bacteria or 20 ng/ul DNA direct from sputum or from bacteria mixed with human DNA was amplified by PCR using published MLST primers for Sa and Pa (45, 46) with Illumina adapters on the 5' ends to enable next generation sequencing of MLST loci (Table S3). PCR amplification of each of the seven MLST loci is performed in triplicate to reduce chances of random PCR bias using reagents listed in Table S4. Triplicate reactions were pooled after PCR and amplified DNA was visualized by agarose gel electrophoresis and quantified by Pico green (Thermo Fisher). After cleaning with Ampure beads (Beckman Coulter), the seven MLST loci for each sample were pooled in equimolar amounts and barcoded with Illumina Nextera XT indexes. PCR amplification, indexing, and cleanup was performed as described in the 16S Metagenomic Sequencing Library Preparation guide (Illumina). Barcoded MLST loci were sequenced on the Illumina MiSeq to produce 2 x 300 bp paired-end reads.

Bioinformatic analysis. Methods outlined below for PopMLST will be available at https://github.com/marade/PopMLST upon publication. Reads were deconvolved based on their locus-specific primer sequence using Python tre, with approximate matching to MLST loci allowing for up to a 25% mismatch (https://github.com/laurikari/tre/). The 3' end of read 2 was trimmed using a binary search algorithm designed to maximize the number of merged reads of the correct size. Trimmed reads for all MLST loci, except *yqi*, were merged using VSEARCH 2.13.4 fastq_mergepairs. Two basepairs of the sequence in the *yqi* locus beyond the 3' ends of

read 1 and 2 (due to the length of this amplicon) were artificially supplied (these bases are conserved according to the MLST database (47)). *yqi* reads were joined using VSEARCH 2.13.4 fastq_join. Merged reads, with their adaptors trimmed using Cutadapt 2.3, were then processed using DADA2 to generate amplicon sequence variants (ASVs) for each locus.

To determine the identity and quantify the relative abundance of each MLST locus, the ASVs were queried against the PubMLST database (https://pubmlst.org/saureus/ and https://pubmlst.org/paeruginosa/) (47) for the appropriate species using BLAST+ BLASTN (66). The matching sequence with the highest identity and longest length (less than or equal to the maximum locus length present in the database) was used to label each ASV by locus type, with less than 100% identity matches being marked as potential novel alleles. The resulting output table includes each MLST loci type identified, the ASV, and the number of reads assigned to each type, much like a classic 16S OTU table.

Author Contributions. Conceptualization, S.M., S.D., and P.S; methodology, S.M., S.D., and M.R.; software, M.R.; validation, S.M. and S.D.; investigation, S.M., S.D., S.R., and D.S.; resources, M.A., S.S., P.J., E.M.; writing, S.M., S.D., and P.S.; supervision, PS.; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding. This study was funded by the Cystic Fibrosis Foundation (SINGH19G0), University of Washington CF Research Development programs (SINGH19R0 and P30DK089507) and NIH (R01HL141098 and K24HL102246) to PKS. Some sample collection was funded by an investigator-initiated grant from Vertex, Inc.

Acknowledgments. The authors would like to thank Hillary Hayden and Michael Parkins for critical review of the manuscript. The authors would like to thank the laboratories of Matthew Parsek and Lucas Hoffman for providing strains.

Conflicts of Interest. The authors declare no conflict of interest

References

1. Gjodsbol K, Christensen JJ, Karlsmark T, Jorgensen B, Klein BM, and Krogfelt KA. Multiple bacterial species reside in chronic wounds: a longitudinal study. *Int Wound J.* 2006;3(3):225-31.

2. McGeachie J. Recurrent Infection of the Urinary Tract: Reinfection or Recrudescence? *British Medical Journal.* 1966;1(5493):2.

3. Frederick JB, Abraham. Anaerobic infection of the paranasal sinuses. *New England Journal of Medicine.* 1974;290(3):3.

4. Branger C, Gardye C, and Lambert-Zechovsky N. Persistence of Staphylococcus aureus strains among cystic fibrosis patients over extended periods of time. *J Med Microbiol.* 1996;45(4):294-301.

5. Kahl BC, Duebbers A, Lubritz G, Haeberle J, Koch HG, Ritzerfeld B, Reilly M, Harms E, Proctor RA, Herrmann M, et al. Population dynamics of persistent Staphylococcus aureus isolated from the airways of cystic fibrosis patients during a 6-year prospective study. *J Clin Microbiol.* 2003;41(9):4424-7.

6. Hirschhausen N, Block D, Bianconi I, Bragonzi A, Birtel J, Lee JC, Dubbers A, Kuster P, Kahl J, Peters G, et al. Extended Staphylococcus aureus persistence in cystic fibrosis is associated with bacterial adaptation. *Int J Med Microbiol.* 2013;303(8):685-92.

7. Struelens MJ, Schwam V, Deplano A, and Baran D. Genome macrorestriction analysis of diversity and variability of Pseudomonas aeruginosa strains infecting cystic fibrosis patients. *J Clin Microbiol.* 1993;31(9):2320-6.

8. Burns JL, Gibson RL, McNamara S, Yim D, Emerson J, Rosenfeld M, Hiatt P, McCoy K, Castile R, Smith AL, et al. Longitudinal Assessment of Pseudomonas aeruginosa in Young Children with Cystic Fibrosis. *The Journal of Infectious Diseases.* 2001;183(3):444-52.

9. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, Miller SI, Ramsey BW, Speert DP, Moskowitz SM, et al. Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A.* 2006;103(22):8487-92.

10. Goerke C, Gressinger M, Endler K, Breitkopf C, Wardecki K, Stern M, Wolz C, and Kahl BC. High phenotypic diversity in infecting but not in colonizing Staphylococcus aureus populations. *Environ Microbiol.* 2007;9(12):3134-42.

11. Long DA-OX, Wolter DA-O, Lee M, Precit MA-O, McLean KA-O, Holmes E, Penewit KA-O, Waalkes AA-O, Hoffman LA-O, and Salipante SA-O. Polyclonality, Shared Strains, and Convergent Evolution in Chronic Cystic Fibrosis Staphylococcus aureus Airway Infection. 1535-4970 (Electronic)).

12. Jelsbak L, Johansen HK, Frost AL, Thogersen R, Thomsen LE, Ciofu O, Yang L, Haagensen JA, Hoiby N, and Molin S. Molecular epidemiology and dynamics of Pseudomonas aeruginosa populations in lungs of cystic fibrosis patients. *Infect Immun.* 2007;75(5):2214-24.

13. Jain M, Ramirez D, Seshadri R, Cullina JF, Powers CA, Schulert GS, Bar-Meir M, Sullivan CL, McColley SA, and Hauser AR. Type III secretion phenotypes of Pseudomonas aeruginosa strains change during infection of individuals with cystic fibrosis. *J Clin Microbiol.* 2004;42(11):5229-37.

14. Anthony M, Rose B, Pegler MB, Elkins M, Service H, Thamotharampillai K, Watson J, Robinson M, Bye P, Merlino J, et al. Genetic analysis of Pseudomonas aeruginosa isolates from the sputa of Australian adult cystic fibrosis patients. *J Clin Microbiol.* 2002;40(8):2772-8.

15. Syrmis MW, O'Carroll MR, Sloots TP, Coulter C, Wainwright CE, Bell SC, and Nissen MD. Rapid genotyping of Pseudomonas aeruginosa isolates harboured by adult and paediatric patients with cystic fibrosis using repetitive-element-based PCR assays. *J Med Microbiol.* 2004;53(Pt 11):1089-96.

16. Mahenthiralingam E, Campbell ME, Foster J, Lam JS, and Speert DP. Random amplified polymorphic DNA typing of Pseudomonas aeruginosa isolates recovered from patients with cystic fibrosis. *J Clin Microbiol.* 1996;34(5):1129-35.

17. Oliver A, Canton R, Campo P, Baquero F, and Blazquez J. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. *Science.* 2000;288(5469):1251-4.

18. Hogardt M, Hoboth C, Schmoldt S, Henke C, Bader L, and Heesemann J. Stage-specific adaptation of hypermutable Pseudomonas aeruginosa isolates during chronic pulmonary infection in patients with cystic fibrosis. *J Infect Dis.* 2007;195(1):70-80.

19. Jorth P, Staudinger BJ, Wu X, Hisert KB, Hayden H, Garudathri J, Harding CL, Radey MC, Rezayat A, Bautista G, et al. Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis Lungs. *Cell Host Microbe.* 2015;18(3):307-19.

20. Langhanki L, Berger P, Treffon J, Catania F, Kahl BC, and Mellmann A. In vivo competition and horizontal gene transfer among distinct Staphylococcus aureus lineages as major drivers for adaptational changes during long-term persistence in humans. *BMC Microbiol.* 2018;18(1):152.

21. McCallum SJ, Corkill J, Gallagher M, Ledson MJ, Hart CA, and Walshaw MJ. Superinfection with a transmissible strain of Pseudomonas aeruginosa in adults with cystic fibrosis chronically colonised by P aeruginosa. *Lancet.* 2001;358(9281):558-60.

22. Salunkhe P, Smart CH, Morgan JA, Panagea S, Walshaw MJ, Hart CA, Geffers R, Tummler B, and Winstanley C. A cystic fibrosis epidemic strain of Pseudomonas aeruginosa displays enhanced virulence and antimicrobial resistance. *J Bacteriol.* 2005;187(14):4908-20.

23. Parkins MD, Glezerson BA, Sibley CD, Sibley KA, Duong J, Purighalla S, Mody CH, Workentine ML, Storey DG, Surette MG, et al. Twenty-five-year outbreak of Pseudomonas aeruginosa infecting individuals with cystic fibrosis: identification of the prairie epidemic strain. *J Clin Microbiol.* 2014;52(4):1127-35.

24. Duong J, Booth SC, McCartney NK, Rabin HR, Parkins MD, and Storey DG. Phenotypic and Genotypic Comparison of Epidemic and Non-Epidemic Strains of Pseudomonas aeruginosa from Individuals with Cystic Fibrosis. *PLOS ONE.* 2015;10(11):e0143466.

25. Yagci S, Hascelik G, Dogru D, Ozcelik U, and Sener B. Prevalence and genetic diversity of Staphylococcus aureus small-colony variants in cystic fibrosis patients. *Clin Microbiol Infect.* 2013;19(1):77-84.

26. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, and Kishony R. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet.* 2014;46(1):82-7.

27. Williams D, Evans B, Haldenby S, Walshaw MJ, Brockhurst MA, Winstanley C, and Paterson S. Divergent, coexisting Pseudomonas aeruginosa lineages in chronic cystic fibrosis lung infections. *Am J Respir Crit Care Med.* 2015;191(7):775-85.

28. Allendorf FW, and Lundquist LL. Introduction: Population Biology, Evolution, and Control of Invasive Species. *Conservation Biology.* 2003;17(1):24-30.

29. Frederiksen B, Koch C, and Hoiby N. Antibiotic treatment of initial colonization with Pseudomonas aeruginosa postpones chronic infection and prevents deterioration of pulmonary function in cystic fibrosis. *Pediatr Pulmonol.* 1997;23(5):330-5.

30. Hansen CR, Pressler T, and Hoiby N. Early aggressive eradication therapy for intermittent Pseudomonas aeruginosa airway colonization in cystic fibrosis patients: 15 years experience. *J Cyst Fibros.* 2008;7(6):523-30.

31. Ratjen F, Doring G, and Nikolaizik WH. Effect of inhaled tobramycin on early Pseudomonas aeruginosa colonisation in patients with cystic fibrosis. *Lancet.* 2001;358(9286):983-4.

32. Treggiari MM, Rosenfeld M, Retsch-Bogart G, Gibson R, and Ramsey B. Approach to eradication of initial Pseudomonas aeruginosa infection in children with cystic fibrosis. *Pediatr Pulmonol.* 2007;42(9):751-6.

33. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, and Swaminathan B. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol.* 1995;33(9):2233-9.

34. van Belkum A, Melles DC, Nouwen J, van Leeuwen WB, van Wamel W, Vos MC, Wertheim HF, and Verbrugh HA. Co-evolutionary aspects of human colonisation and infection by Staphylococcus aureus. *Infect Genet Evol.* 2009;9(1):32-47.

35. Kidd TJ, Grimwood K, Ramsay KA, Rainey PB, and Bell SC. Comparison of three molecular techniques for typing Pseudomonas aeruginosa isolates in sputum samples from patients with cystic fibrosis. *J Clin Microbiol.* 2011;49(1):263-8.

36. Waters V, Zlosnik JEA, Yau YCW, Speert DP, Aaron SD, and Guttman DS. Comparison of three typing methods for Pseudomonas aeruginosa isolates from patients with cystic fibrosis. *European Journal of Clinical Microbiology & Infectious Diseases.* 2012;31(12):3341-50.

37. Gentili V, Gianesini S, Balboni PG, Menegatti E, Rotola A, Zuolo M, Caselli E, Zamboni P, and Di Luca D. Panbacterial real-time PCR to evaluate bacterial burden in chronic wounds treated with Cutimed™ Sorbact™. *Eur J Clin Microbiol Infect Dis.* 2012;31(7):1523-9.

38. Yusuf E, Jordan X, Clauss M, Borens O, Mäder M, and Trampuz A. High bacterial load in negative pressure wound therapy (NPWT) foams used in the treatment of chronic wounds. *Wound Repair and Regeneration.* 2013;21(5):677-81.

39. Rudkjøbing VB, Aanaes K, Wolff TY, von Buchwald C, Johansen HK, and Thomsen TR. An exploratory study of microbial diversity in sinus infections of cystic fibrosis patients by molecular methods. *Journal of Cystic Fibrosis.* 2014;13(6):645-52.

40. Fischer AJ, Singh SB, LaMarche MM, Maakestad LJ, Kienenberger ZE, Peña TA, Stoltz DA, and Limoli DH. Sustained Coinfections with Staphylococcus aureus and Pseudomonas aeruginosa in Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine.* 2021;203(3):328-38.

41. Hisert KB, Heltshe SL, Pope C, Jorth P, Wu X, Edwards RM, Radey M, Accurso FJ, Wolter DJ, Cooke G, et al. Restoring Cystic Fibrosis Transmembrane Conductance Regulator Function Reduces Airway Bacteria and Inflammation in People with Cystic Fibrosis and Chronic Lung Infections. *Am J Respir Crit Care Med.* 2017;195(12):1617-28.

42. Shevchenko SG, Radey M, Tchesnokova V, Kisiela D, and Sokurenko EV. Escherichia coli Clonobiome: Assessing the Strain Diversity in Feces and Urine by Deep Amplicon Sequencing. *Appl Environ Microbiol.* 2019;85(23).

43. Rosen MJ, Davison M, Bhaya D, and Fisher DS. Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science.* 2015;348(6238):1019-23.

44. Gan GL, Willie E, Chauve C, and Chindelevitch L. Deconvoluting the diversity of within-host pathogen strains in a multi-locus sequence typing framework. *BMC Bioinformatics.* 2019;20(Suppl 20):637.

45. Curran B, Jonas D, Grundmann H, Pitt T, and Dowson CG. Development of a multilocus sequence typing scheme for the opportunistic pathogen Pseudomonas aeruginosa. *J Clin Microbiol.* 2004;42(12):5644-9.

46. Enright MC, Day NP, Davies CE, Peacock SJ, and Spratt BG. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J Clin Microbiol.* 2000;38(3):1008-15.

47. Jolley KA, Bray JE, and Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 2018;3(124.

48. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, and Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13(7):581-3.

49. Polz MF, and Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol.* 1998;64(10):3724-30.

50. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 2011;39(13):e90.

51. Schirmer M, D'Amore R, Ijaz UZ, Hall N, and Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics.* 2016;17(125.

52. McAdam PR, Holmes A, Templeton KE, and Fitzgerald JR. Adaptive evolution of Staphylococcus aureus during chronic endobronchial infection of a cystic fibrosis patient. *PLoS One.* 2011;6(9):e24301.

53. Morgan S, Lippman S, Bautista G, Harrison JJ, Harding CL, Gallagher L, Cheng A, Siehnel R, Ravishankar S, Usui M, et al. Bacterial Fitness in Chronic Wound Infections is Primarily Mediated by the Capacity For High-Density Growth, Not Virulence or Biofilm Functions. *PLoS Pathog.*

54. Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, Xue K, Liu F, Deng Y, Liang Y, et al. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLOS ONE.* 2017;12(4):e0176716.

55. Feigelman R, Kahlert CR, Baty F, Rassouli F, Kleiner RL, Kohler P, Brutsche MH, and von Mering C. Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome.* 2017;5(1):20.

56. Wolcott RD, Hanson JD, Rees EJ, Koenig LD, Phillips CD, Wolcott RA, Cox SB, and White JS. Analysis of the chronic wound microbiota of 2,963 patients by 16S rDNA pyrosequencing. *Wound Repair Regen.* 2016;24(1):163-74.

57. Loesche M, Gardner SE, Kalan L, Horwinski J, Zheng Q, Hodkinson BP, Tyldsley AS, Franciscus CL, Hillis SL, Mehta S, et al. Temporal Stability in Chronic Wound Microbiota Is Associated With Poor Healing. *J Invest Dermatol.* 2017;137(1):237-44.

58. Price LB, Liu CM, Melendez JH, Frankel YM, Engelthaler D, Aziz M, Bowers J, Rattray R, Ravel J, Kingsley C, et al. Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PLoS One.* 2009;4(7):e6462.

59. Stokell JR, Gharaibeh RZ, Hamp TJ, Zapata MJ, Fodor AA, and Steck TR. Analysis of changes in diversity and abundance of the microbial community in a cystic fibrosis patient over a multiyear period. *J Clin Microbiol.* 2015;53(1):237-47.

60. Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, Cavalcoli JD, VanDevanter DR, Murray S, Li JZ, et al. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc Natl Acad Sci U S A.* 2012;109(15):5809-14.

61. Klepac-Ceraj V, Lemon KP, Martin TR, Allgaier M, Kembel SW, Knapp AA, Lory S, Brodie EL, Lynch SV, Bohannan BJ, et al. Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and Pseudomonas aeruginosa. *Environ Microbiol.* 2010;12(5):1293-303.

62. Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, Daly RA, Karaoz U, Andersen GL, Brown R, Fujimura KE, et al. Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients. *PLoS One.* 2010;5(6):e11044.

63. Coburn B, Wang PW, Diaz Caballero J, Clark ST, Brahma V, Donaldson S, Zhang Y, Surendra A, Gong Y, Elizabeth Tullis D, et al. Lung microbiota across age and disease stage in cystic fibrosis. *Sci Rep.* 2015;5(10241.

64. Whiteside SA, Razvi H, Dave S, Reid G, and Burton JP. The microbiome of the urinary tract--a role beyond infection. *Nat Rev Urol.* 2015;12(2):81-90.

65. McCallum SJ, Gallagher MJ, Corkill JE, Hart CA, Ledson MJ, and Walshaw MJ. Spread of an epidemic Pseudomonas aeruginosa strain from a patient with cystic fibrosis (CF) to non-CF relatives. *Thorax.* 2002;57(6):559.

66. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10(1):421.

Table 1. PopMLST correctly identifies single *S. aureus* and *P. aeruginosa* isolates as a single MLST type.

Sample	Percent of reads mapping to single loci type (SEM)
S. aureus	
NCTC8325	100% (0)
NCTC8325 (repeat)	100% (0)
MN8	100% (0)
MN8 (Repeat)	100% (0)
Newman	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
P. aeruginosa	
PA14	100% (0)
PA14 (repeat)	99.96% (0.0272)*
LES	100% (0)
LES (repeat)	100% (0)
PAO1	100% (0)
PAO1 (repeat)	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	100% (0)
CF isolate	99.6% (0.006)**
CF isolate	100% (0)
Wound isolate	100% (0)
Wound isolate	100% (0)

* Three of the six loci indicated the presence of a second loci type at less than 1% likely due to sequencing error.
* Four of the six loci indicated the presence of a second loci type at less than 1% likely due to sequencing error.

Table 2. Frequency of false positive calls depending on criteria used to call MLST type presence. Accuracy is shown as function of the abundance and number of loci used to make calls.

Abundance of loci used to call MLST type presence	*P. aeruginosa*		*S. aureus*	
	1 locus used to call MLST type presence	2 loci used to call MLST type presence	1 locus used to call MLST type presence	2 loci used to call MLST type presence
	# experiments with false positive (%)	# experiments with false positive (%)	# experiments with false positive (%))	# experiments with false positive (%)
5%	0/38 (0%)	0/38 (0%)	0/41 (0%)	0/41 (0%)
4%	0/38 (0%)	0/38 (0%)	0/41 (0%)	0/41 (0%)
3%	0/38 (0%)	0/38 (0%)	2/41 (5%)	0/41 (0%)
2%	3/38 (8%)	0/38 (0%)	5/41 (12%)	0/41 (0%)
1%	4/38 (11%)	0/38 (0%)	7/41 (17%)	0/41 (0%)
0.50%	5/38 (13%)	1/38 (3%)	7/41 (17%)	0/41 (0%)
0.25%	6/38 (16%)	1/38 (3%)	8/41 (20%)	0/41 (0%)
0.10%	9/38 (23%)	2/38 (5%)	10/41 (24%)	2/41 (5%)
0.01%	14/38 (37%)	7/38 (18%)	10/41 (24%)	2/41 (5%)
0%	14/38 (37%)	7/38 (18%)	10/41 (24%)	2/41 (5%)

Table 3. PopMLST's sensitivity for detecting known *S. aureus* MLST types

% low abundance MLST type	MLST Loci: arc # times loci detected (range)	aro # times loci detected (range)	glp # times loci detected (range)	gmk # times loci detected (range)	pta # times loci detected (range)	tpi # times loci detected (range)	yqi # times loci detected (range)
10%	4/4 (19.5-2.8)	4/4 (6.0-1.8)	4/4 (10.6-1.9)	4/4 (14.0-2.8)	4/4 (14.1-2.3)	4/4 (14.9-2.4)	4/4 (16.0-2.3)
5%	4/4 (5.3-3.8)	4/4 (3.2-1.6)	4/4 (6.1-1.8)	4/4 (5.2-2.7)	4/4 (4.6-3.0)	4/4 (5.5-3.4)	4/4 (4.7-3.9)
2%	4/4 (2.4-1.7)	4/4 (1.2-0.6)	4/4 (3.4-0.7)	4/4 (11.1-2.0)	3/4 (1.7-0)	4/4 (2.3-1.4)	3/4 (1.9-0)
1%	3/4 (2.2-0)	4/4 (1.1-0.1)	3/4 (1.5-0)	4/4 (1.5-0.2)	2/4 (1.4-0)	3/4 (1.5-0)	3/4 (2.0-0)
0.1%	2/3 (0.4-0)	3/3 (0.3-0.1)	2/3 (0.3-0)	3/3 (1.5-0.2)	0/3 (0-0)	3/3 (0.3-0.2)	1/3 (0.2-0)

Table 4. PopMLST's sensitivity for detecting known *P. aeruginosa* MLST types

% low abundance MLST type	MLST loci: acs # times loci detected (range)		gua* # times loci detected (range)		mut # times loci detected (range)		nuo* # times loci detected (range)		pps # times loci detected (range)		trp # times loci detected (range)	
10%	4/4	(13.5-6.2)	3/3	(24.8-4.1)	4/4	(14.9-6.9)	3/3	(13.3-8.2)	4/4	(14.8-3.9)	4/4	(15.2-6.0)
5%	3/3	(7.0-4.6)	2/2	(6.6-4.8)	3/3	(7.7-2.4)	1/2	(6.6-1)	3/3	(7.6-4.6)	3/3	(7.7-2.8)
2%	3/3	(3.18-1.7)	2/2	(2.3-2.1)	3/3	(3.2-1.4)	2/2	(2.8-1.9)	3/3	(4.2-1.7)	3/3	(3.5-2.0)
1%	4/4	(1.6-0.7)	3/3	(4.0-0.4)	4/4	(2.1-0.7)	3/3	(1.6-1.1)	4/4	(2.7-0.5)	4/4	(1.8-0.7)
0.1%	3/3	(0.3-0.1)	2/3	(0.5-0)	3/3	(0.1-0.06)	1/3	(0.4-0)	3/3	(0.3-0.08)	3/3	(0.2-0.08)

***** Two of the strains used for one of the 4 mixing experiments shared alleles at these loci, because the two strains could not be differentiated at these loci, they were eliminated from analysis for these loci.

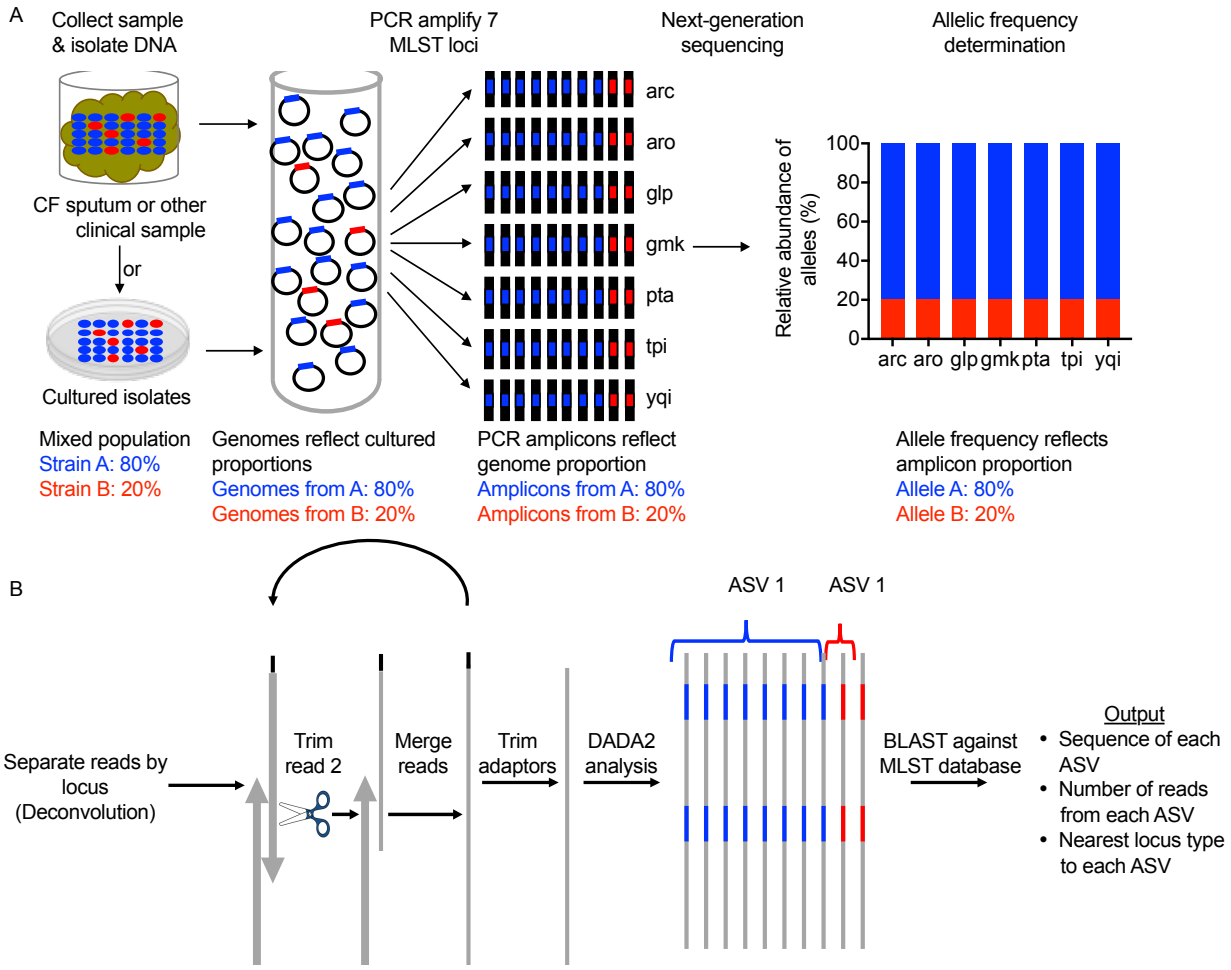

Figure 1. PopMLST methods. (A) PopMLST can be performed on clinical specimens (without culturing) or cultured isolates. MLST loci are PCR amplified in separate reactions, amplicons Illumina-sequenced, and the relative abundance of reads representing MLST loci are measured. (B) Bioinformatic analysis deconvolutes reads using permissive alignment to assign them to MLST loci, tests read 2 trimming lengths to optimize merging, removes adaptors, identifies amplicon sequence variants (ASVs) using DADA2, and uses BLAST to identify closest MLST locus type.

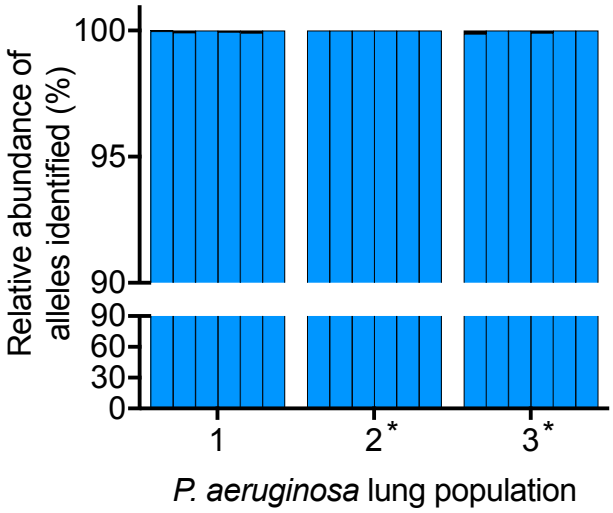

Figure 2. PopMLST correctly identifies genetically-diversified clonally-related Pa as a single MLST type. Plot shows the relative abundance of each MLST allele that matches the known MLST sequence (determined by genome sequencing). The six bars for each sample show the relative abundance of *acs, gua, mut, nuo, pps,* and *trp* loci (in order). Black bars indicate any additional MLST loci types detected and in all cases were less than 0.2%. * indicates hypermutable populations due to *mutS* (population 2) and *mutL* (population 3) mutations.

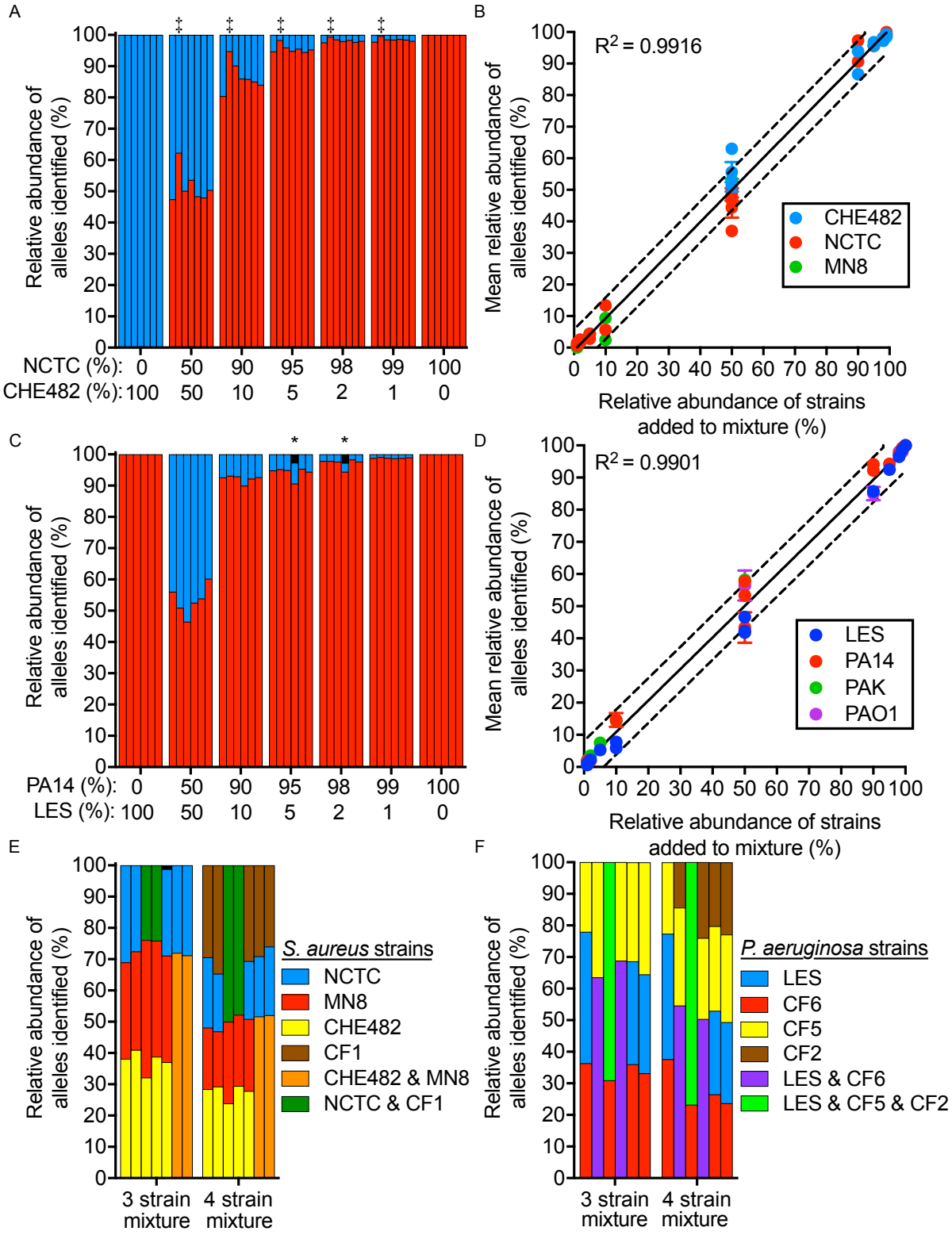

Figure 3. PopMLST measures strain relative abundance. (**A and C**) DNA from two Sa (**A**) and Pa (**C**) strains with different MLST types were mixed at indicated ratios. Red bars indicate reads that PopMLST called as alleles from Sa NCTC8325 (A) and Pa PA14 (C); blue bars

indicate reads called as alleles from Sa CHE482 (A) and Pa LES (C). (**B and D**) shows the mean and SEM of reads mapping to indicated MLST type from 21 independent two-strain Sa mixtures (**B**) including NCTC8325 (red) with CHE482 (blue) or MN8 (green); and 20 independent two-strain Pa mixtures (**D**) of PA14 (red) and LES (blue) or PAK (green), or PAO1 (purple) and LES (blue). Data for individual alleles can be found in Figure S2 and S3. Some error bars (SEM) were smaller than symbols; solid line indicates expected result, dashed lines indicate +/- 10%. (**E and F**) Unique MLST loci alleles of the four strains tested are shown as blue, red, yellow, and brown. Alleles shared between two strains are colored as follows (red & blue = purple; yellow & red = orange; blue & brown = green (E) or blue & yellow & brown = green (F)). Bars in (A and E) show relative abundance of *arc, aro, glp, gmk, pta, tpi,* and *yqi* (in order). Bars in (C and F) show relative abundance of *acs, gua, mut, nuo, pps,* and *trp* (in order). MLST alleles identified but not present in the mixtures (likely sequencing error), are indicated in black and those detected at >1%, are indicated with *. ‡ indicates PCR bias as evidenced by one allele being consistently under or overrepresented.

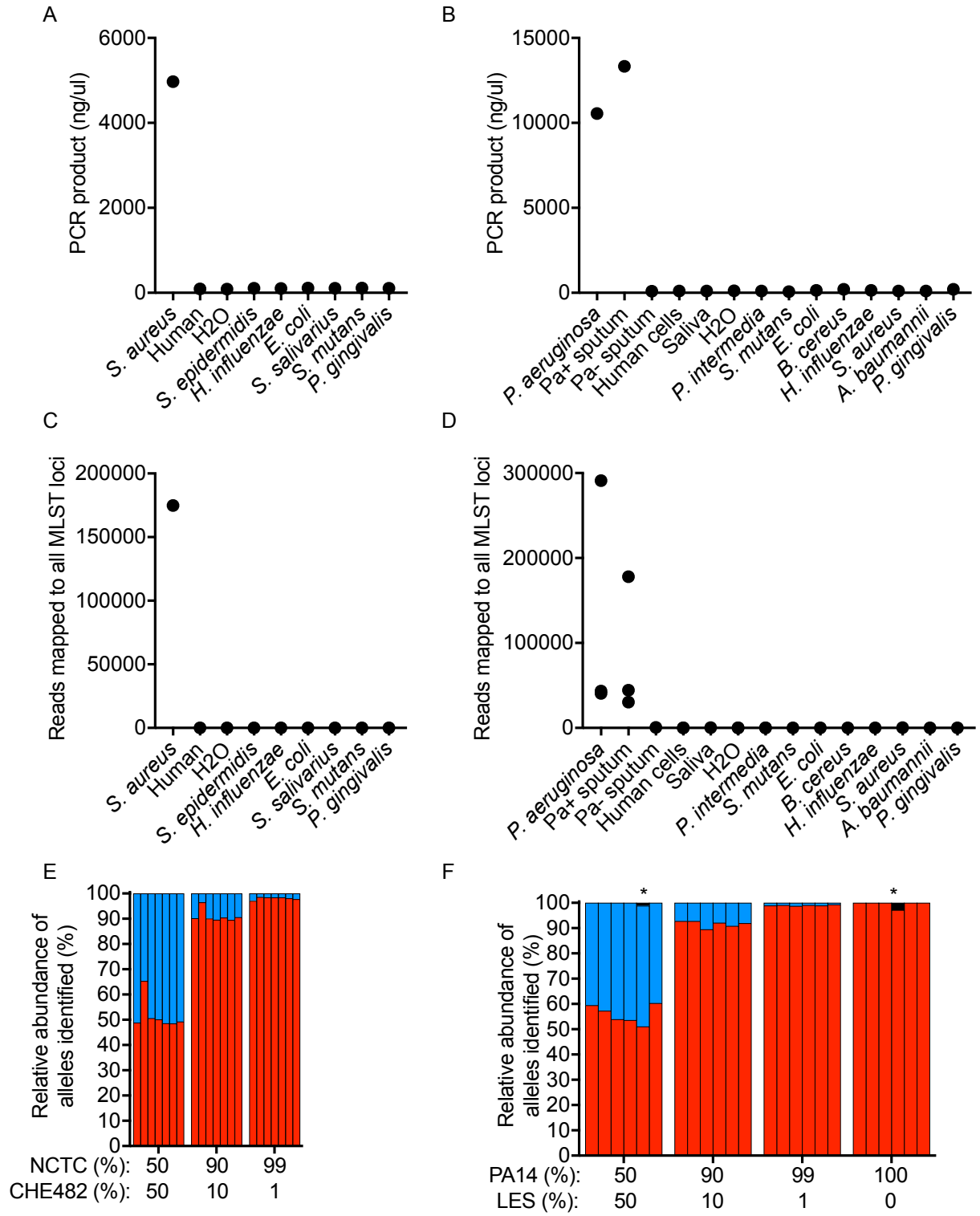

Figure 4. Heterologous DNA does not interfere with PopMLST. Concentrations of product after PCR using PopMLST primers for Sa (**A**) and Pa (**B**) on DNA from the indicated sources; 'human' indicates DNA extracted from tissue culture cells; 'H2O' indicates ultrapure water; 'Pa+ sputum' indicates sputum from a CF subject culture-positive for Pa; and 'Pa- sputum'

indicates sputum from a CF subject culture-negative for Pa. The sum of reads produced by PopMLST that mapped to seven Sa (**C**) or six Pa (**D**) MLST loci are shown for samples containing target and non-target DNA from PCR reactions in A and B. (**E and F**) 95% human DNA from tissue culture cells was added to the same mixtures of two control strains from Figure 2C. (**E**) Bars for each mixture show relative abundance *arc, aro, glp, gmk, pta, tpi,* and *yqi* matching the MLST type of NCTC8325 (red) or CHE482 (blue). (**F**) Bars for each mixture show relative abundance of *acs, gua, mut, nuo, pps,* and *trp* (in order) matching the MLST type of PA14 (red) or LES (blue). * indicates the presence of an unexpected loci type (black), likely due to sequencing error.

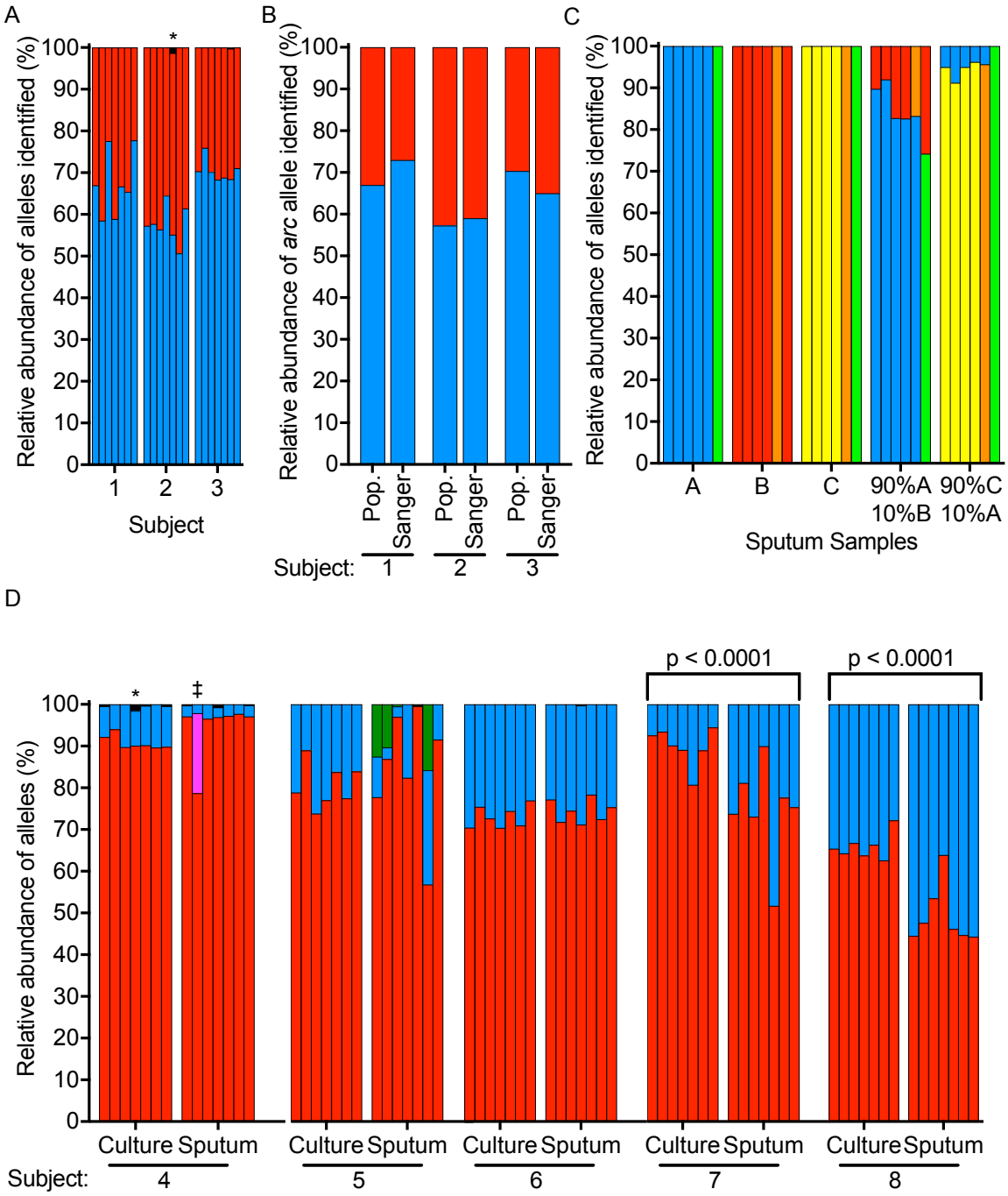

Figure 5. PopMLST measures MLST types in clinical samples. **A**. PopMLST was performed on ~95 Sa isolates cultured from 3 CF subjects, blue and red bars indicate different MLST loci alleles. The seven bars for each sample show relative abundance of *arc, aro, glp, gmk, pta, tpi,* and *yqi* (in order). The *pta* locus showed a third allele (black bar) which was 1 nucleotide different than the predominant allele and likely represents sequencing error. **B**. The relative abundance of the *arc* locus as measured by PopMLST (Pop) and by individually Sanger sequencing (Sanger) the *arc* locus of 20-30 isolates from each sample represented in A. **C**. PopMLST performed directly on DNA isolated from sputum (samples A, B, C) and from

indicated mixtures of these samples. Red, blue, and yellow bars indicate the abundance of MLST alleles corresponding to sample A, B, and C respectively, with green bars indicating an allele shared between A and C and orange bars indicating an allele shared between B and C. Control experiments examining >100 Pa isolates cultured from sputum A, B, and C showed each contained a single Pa MLST type. **D**. Sa PopMLST of scrapes of >100 colonies (culture) and directly from sputum (sputum). Red and blue indicate allele frequencies of the predominant and secondary MLST types, respectively. Sputum of subject 5 contained three loci from an additional allele, likely indicating a third un-cultured MLST type (green). Presence of the MLST types shown in red and blue were confirmed by MLST typing of single isolates. Significant difference ($p < 0.01$, by multiple t-test) in abundance of the minor allele across the 7 loci is indicated by the p-value. ‡ indicates detection of a third loci allele with a single SNP different than the predominant loci, likely due to either sequencing error or mutation; * indicates the presence of a third allele (black), likely due to sequencing error.