Pan-PCR, a Computational Method for Designing Bacterium-Typing Assays Based on Whole-Genome Sequence Data

Joy Y. Yang,a Shelise Brooks,b Jennifer A. Meyer,a Robert R. Blakesley,b Adrian M. Zelazny,c Julia A. Segre,a Evan S. Snitkin

Genetics and Molecular Biology Branch, National Human Genome Research Institute, Bethesda, Maryland, USAa; NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland, USAa; Department of Laboratory Medicine, National Institutes of Health Clinical Center, National Institutes of Health, Bethesda, Maryland, USAc

With increasing rates of antibiotic resistance, bacterial infections have become more difficult to treat, elevating the importance of surveillance and prevention. Effective surveillance relies on the availability of rapid, cost-effective, and informative typing methods to monitor bacterial isolates. PCR-based typing assays are fast and inexpensive, but their utility is limited by the lack of targets which are capable of distinguishing between strains within a species. To identify highly informative PCR targets from the growing database of publicly available bacterial genome sequences, we developed pan-PCR. This computer algorithm uses existing genome sequences for isolates of a species of interest and identifies a set of genes whose patterns of presence or absence provide the best discrimination between strains in this species. A set of PCR primers targeting the identified genes is then designed, with each PCR product being of a different size to allow multiplexing. These target DNA regions and PCR primers can then be utilized to type bacterial isolates. To evaluate pan-PCR, we designed an assay for the emerging pathogen Acinetobacter baumannii. Taking as input a set of 29 previously sequenced genomes, pan-PCR identified 6 genetic loci whose presence or absence was capable of distinguishing all the input strains. This assay was applied to a set of patient isolates, and its discriminatory power was compared to that of multilocus sequence typing (MLST) and whole-genome optical maps. We found that the pan-PCR assay was capable of making clinically relevant distinctions between strains with identical MLST profiles and showed a discriminatory power similar to that of optical maps. Pan-PCR represents a tool capable of exploiting available genome sequence data to design highly discriminatory PCR assays. The ease of design and implementation makes this approach feasible for diagnostic facilities of all sizes.

Hospital environments are host to a diverse population of potentially infectious agents (1). Approximately 5% of patients admitted to acute care hospitals acquire at least one infection. In 2002, the CDC estimated that there were 1.7 million hospital-acquired infections (HAIs), with approximately 99,000 associated deaths (2).

With the rising incidence of antibiotic resistance, bacterial infections have become increasingly difficult to treat, heightening the need for effective surveillance and infection control practices to limit the impact of HAIs. Upon isolation of multidrug-resistant organisms, a critical task for the clinical microbiology laboratory is assessing whether the strain has been previously seen in that institution. Early distinction between nosocomial transmission events versus importation of new strains allows a targeted response by infection control teams toward prevention of spread of the organism (3).

An effective strain-typing methodology is required to distinguish between intra- and interhospital spread. Fine-resolution typing methods often rely upon differences in DNA nucleotide sequence, gene content, and/or genome architecture. The ideal typing approach should be rapid, inexpensive, capable of distinguishing between even closely related strains of the same species, and capable of comparing contemporary isolates to previously typed isolates using computational analysis. Current bacterium-typing methods such as pulsed-field gel electrophoresis (PFGE), multilocus sequence typing (MLST) (4, 5), and generalized PCR-based methods come with trade-offs in speed, cost, and resolution. PFGE yields a high degree of resolution, and the required equipment is now found in many hospital labs; however, it requires several days to run and requires specialized training. The resolution provided by MLST is comparable to that of PFGE, but since it requires sequencing, it can also take a few days and is relatively expensive. PCR-based methods can be less expensive and require only a few hours; however, currently, generalized methods such as repetitive sequence-based PCR (Rep-PCR) (6) and randomly amplified polymorphic DNA PCR (RAPD-PCR) (7) or arbitrarily primed PCR (AP-PCR) (8) typically yield lower degrees of resolution than PFGE or MLST.

The decreasing cost and increasing speed of whole-genome sequencing have led some to posit that whole-genome sequencing may become the standard for bacterial typing in the future (9, 10). Several groups, including our own, have published studies employing whole-genome sequencing to investigate hospital outbreaks in clinically relevant turnaround times (11, 12). However, in the near term, routine use of whole-genome sequencing remains limited by cost and the lack of bioinformatics tools to easily translate these data into succinct strain-typing information. Despite these barriers to the widespread adoption of diagnostic
genome sequencing, recent work has demonstrated that the knowledge gained by whole-genome sequencing has the potential to spur the development of new molecular diagnostics. Several groups are attempting to take advantage of existing genome sequence data to design sequence- and PCR-based typing assays. In the last year, PCR-based typing assays have been published for organisms such as *Streptococcus pneumoniae* (13), *Neisseria meningitidis* (14), and *Acinetobacter baumannii* (15).

The design of PCR typing assays from genome sequence data exploits the concept of the bacterial species pan-genome (16), a term referring to the observed variation in gene content of strains within a given species (17). Variation in gene content occurs during the evolution of many bacterial species and provides the opportunity to distinguish different strains by probing for the presence or absence of these variable regions. Essential to the design of an effective assay is the ability to identify a set of genes whose variable presence is capable of distinguishing strains of interest. One approach that has been taken to automate this process is to computationally find signatures that exist within a group of target genomes but that do not exist in other background genomes (18, 19). A limitation of this approach is that it can determine only whether a given isolate is a member of a particular strain and cannot discern relationships among multiple strains. An extension of this approach identifying multiple distinguishing targets was recently proposed (20), but its clinical utility was limited by requiring the user to design the clinical assay and by not considering the functions of target regions.

Here, we present pan-PCR, which takes genome sequences for a set of strains of interest as the input and returns primers for a multiplex PCR assay capable of distinguishing the strains of interest. As a proof of principle, we apply our approach to design a typing scheme for the nosocomial pathogen *Acinetobacter baumannii* and demonstrate that this assay can identify clinically relevant relationships among patient isolates.

**MATERIALS AND METHODS**

**Data preparation.** Pan-PCR takes a set of genome sequences, including the annotation of their protein-encoding genes, as its input. The nucleotide sequences for coding regions are pooled, and the entire set is clustered by the software program CD-Hit (21) (freely available online), according to sequence identity. The CD-Hit default parameters for pan-PCR are as follows: (i) cluster sequences at 90% sequence identity, using a word size of 8; (ii) group sequences together with their best match rather than the first match within the 90% threshold; and (iii) require that the 100% alignment cover the full length of the shortest sequence. Gene clusters annotated with phrases indicative of elements that may evolve more rapidly than the rest of the genome are discarded. These phrases include “phage,” “transpos,” “frame” and “shift,” “integrase,” and “hypothetical.” Genomes with <5% difference in present/absent gene cluster profiles are collapsed down to one representative. This step prevents selection bias of gene clusters that differentiate only between closely related strains over-represented in the initial input set. This final set of gene clusters and genomes is then passed on to the gene-picking algorithm (Fig. 1).

**PCR target selection algorithm.** The minimum number of PCR targets needed depends on the number of strains (*N*) in the input set. Because the presence/absence of PCR targets is a binary code, the theoretical minimum number of PCR targets needed to completely distinguish a set of strains from each other is \( \log_2 N \), so that 4 strains can be distinguished with a minimum of 2 PCR targets and 8 strains can be distinguished with a minimum of 3 PCR targets (see Fig. S1 in the supplemental material). However, this lower bound is not always achievable, as input strains will often not contain gene presence/absence values to form the shortest binary encoding. Finding the minimum possible set of targets to distinguish all input genomes is an intractable problem, with no polynomial time algorithm that can find the exact solution being available (22). Because of this, an exact algorithm will not be of practical use as the number of available genomes increases. Therefore, we employed a greedy approximation algorithm that iteratively selects gene clusters by maximizing the additional number of strain pairs distinguishable from each other, as described by Laing et al. (20). Specifically, let \( x \) represent a gene profile over the strains, \( k \) represent a unique strain profile with gene profile \( x \) included, \( n_k \) represent the number of times that \( k \) appears over the strains, and \( N \) represent the total number of strains. Then, at each iteration, a gene cluster, \( y \), is picked such that

\[
y = \arg\max_x \left\{ \sum_{k=1}^N \frac{n_k}{2} - \sum_{k=1}^N \frac{n_k^2}{2} \right\}
\]

A primer is then designed to sensitively and specifically amplify this gene...
cluster. If such a primer cannot be found, then the iteration is repeated with this gene cluster thrown out of the analysis. The main step of the algorithm was written in Java, with primer picking and testing calls performed on the primer3 (23) and BLAST (24) programs, respectively, and all steps were integrated using the R package. The software packages used by this program are all freely available.

Genomic data set. To design PCR typing assays for A. baumannii with pan-PCR, a set of 29 A. baumannii genomes was downloaded from the Pathosystems Resource Integration Center (PATRIC) (25) (downloaded on 27 October 2011; see Fig. S2 in the supplemental material for the list of genomes). These 29 genomes, consisting of a total of 108,663 genes, were used as the reference set from which the gene clusters and primers were picked. For validation of the selected primers and comparison of the assay to MLST, an additional set of 7 previously sequenced A. baumannii genomes derived from isolates at the National Institutes of Health Clinical Center (NIHCC) was used (Fig. 2).

PCR conditions. Criteria used for the picking of primers can be altered by the user; however, for the best results, adherence to the default parameters is suggested, constraining common features such as melting temperature and minimum G+C content. In order to obtain PCR product bands of about the same intensity, the molar amount of primers used was scaled to be inversely proportional to the predicted product length (see Table 1 for specific lengths). Fifty to 150 ng of purified DNA was used as the input template. A Bio-Rad DNA Engine Tetrad 2 Peltier thermal cycler was programmed for an initial denaturation at 95°C for 5 min, followed by 20 cycles of denaturation (95°C for 30 s), annealing (60°C for 30 s), and extension (72°C for 1 min 30 s) and a single final extension at 72°C for 7 min. The enzyme mix used for the reaction was TaKaRa LA Taq. The PCR products were run on a 1% agarose gel.

Interpretation. As with any classification method based on a discrete number of outputs (e.g., gel banding patterns), the question of how to assess the degree of relatedness among strains arises. Here, differences among strains typed with pan-PCR were quantified by simply counting the number of band differences. Sets of strains were then compared as a group by applying hierarchical clustering to the differences in banding patterns. For comparison to MLST and optical mapping, the number of band differences between each strain was compared to the number of MLST locus differences and optical map distances, respectively.

Comparison with other methods. We compared the resolution of the pan-PCR assay to that of MLST (5,26) using the set of 29 input genomes as well as the set of 7 sequenced genomes. MLST profiles were found using BLAST against the Institut Pasteur A. baumannii MLST genes (4). Pan-PCR results were also assessed in comparison to optical map pairwise distances calculated using the OpGen platform. Optical maps represent

### Table 1: Output summary of gene clusters and PCR target/primer properties

<table>
<thead>
<tr>
<th>Gene annotation</th>
<th>Primer</th>
<th>Sequence</th>
<th>T_m (°C)</th>
<th>Product size (bp)</th>
<th>Amt (µl) used (20 µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer membrane receptor proteins, mostly Fe transport</td>
<td>1F</td>
<td>CTCCTATTATTCGTTGGGCCAGG</td>
<td>61.3</td>
<td>794</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>1R</td>
<td>CTCCAGGCAGCTATAAGGCTTC</td>
<td>61.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfate permease</td>
<td>2F</td>
<td>GTCACTGTTAAGACGATGG</td>
<td>62.9</td>
<td>612</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>2R</td>
<td>GACCAATCTGAGGACAGCC</td>
<td>59.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-chain fatty acid transport protein</td>
<td>3F</td>
<td>GGAGTACGGGAGCTGATGG</td>
<td>62.5</td>
<td>470</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>3R</td>
<td>AGTGCTCCCATCTTTACCAATG</td>
<td>60.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptional regulator, LysR family</td>
<td>4F</td>
<td>ATAAAGATCGAGTGGCCGCG</td>
<td>60.5</td>
<td>361</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>4R</td>
<td>TCGATTGAACCCATTCTGCGGC</td>
<td>60.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroxymethylglutaryl coenzyme A lyase</td>
<td>5F</td>
<td>GTTAATTTGCTGGAGCAGTTT</td>
<td>60.3</td>
<td>268</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>5R</td>
<td>CATGCGCTAAAGAAGCTGGTAAAG</td>
<td>62.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merops peptidase family S24</td>
<td>6F</td>
<td>CCGATGGGTATATGAGCAGAG</td>
<td>62.9</td>
<td>100</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>6R</td>
<td>AACTGCGACTAAAGGCACCTTC</td>
<td>61.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

_T_m, melting temperature._

![FIG 2 Genomes in the test set. The numbered marks denote the location of the amplicons designed by pan-PCR (refer to Table 1), and the gene name marks refer to Institut Pasteur MLST positions. A, A-nonMDR (A_nonMDR); B, C, and D are A. baumannii isolates identified at our institution (16). ATCC 19606 and ACICU are previously sequenced strains of A. baumannii.](http://jcm.asm.org/issue/03/03/754)
ordered restriction maps of each strain, which amounts to a more informative and digital equivalent of the type of genomic insight provided by pulsed-field gel electrophoresis (27). These optical maps were generated with the NcoI restriction enzyme as described previously (28).

**RESULTS**

**Pan-PCR.** Pan-PCR is an algorithm that utilizes existing publicly available whole-genome bacterial sequence data to automatically design multiplex PCR typing assays. The user selects a set of genomes from within a species with variable gene content, and the computer program outputs a nearly minimal set of PCR primers capable of distinguishing among the input strains (Fig. 1). This is accomplished by identifying a set of genes whose pattern of presence and absence among the input genomes represents a unique signature for each input genome. Primers producing a different-sized PCR product are designed for each targeted gene, such that the assay can be run in a multiplex fashion.

**Design of assay for Acinetobacter baumannii.** To test the validity of this approach, pan-PCR was applied to the nosocomial pathogen *Acinetobacter baumannii*. *Acinetobacter baumannii* has become a formidable clinical pathogen, developing various mechanisms of resistance, with reports of some strains being resistant to all known antibiotics (29). *A. baumannii* is prone to nosocomial transmission due to environmental stability and its ability to asymptptomatically colonize individuals (30), thereby allowing patients to act as undetected transmission sources (31). *A. baumannii* can be difficult to type because of high clonality within major hospital-associated lineages (32–35). Further complicating typing is the recent observation of widespread recombination in clinical isolates of *A. baumannii*, which may lead to MLST and PFGE yielding contradicting results (16). Previous genomic analyses of *A. baumannii* have shown that even among closely related strains, there can be extensive variation in gene content. Thus, we hypothesized that typing of *A. baumannii* on the basis of gene content might achieve resolution beyond that of other molecular typing methods (36).

The input for pan-PCR was 29 strains of *A. baumannii* which had previously been sequenced by various groups and deposited into public databases. The number and the diversity of the input genomes can be varied on the basis of the level of resolution desired. Each *A. baumannii* genome was ~4 Mb with 3,750 genes, on average. From these 29 genomic sequences, the total set of 108,663 genes was first clustered by sequence identity into 9,012 gene clusters. This set of gene clusters was further reduced to 4,376 by removing gene clusters with over 50% of their members annotated as phages, mobile elements, or computer-predicted genes without experimental confirmation (see Materials and Methods and Fig. S3 in the supplemental material). This filtering step was taken to remove genes likely to lack the stability necessary to design primers effective across a diverse set of strains. From the 29 genomes, 22 were chosen as representatives (see Materials and Methods) to avoid biases resulting from redundancies in the available sequence database. Applying our iterative algorithm produced 6 gene clusters (Table 1) whose pattern of presence and absence distinguished all representative genomes *in silico*. The 6 sets of primers probing these gene clusters were validated using a set of 7 sequenced genomes (Fig. 2). The experimental PCR results perfectly matched the *in silico* PCR predictions (Fig. 3), validating both the targets selected and the primers designed by our approach.

**Comparison of resolution of pan-PCR assay with that of MLST.** To assess the level of resolution provided by pan-PCR, we compared our *A. baumannii* assay to other commonly utilized methods. We first performed a comparison with MLST, which is used for typing of diverse bacterial species and has been successfully applied to typing of *A. baumannii* (4, 26, 32). Because MLST classification is based on the sequences of housekeeping genes, it is often unable to distinguish closely related strains. We therefore hypothesized that if the rate of variation in gene content exceeds the evolutionary rate at these housekeeping genes, pan-PCR may effectively distinguish strains that are identical in their MLST regions. In fact, on the input set of 29 genomes, we found that the assay designed by pan-PCR was able to distinguish pairs of genomes with identical MLST profiles, whereas the opposite
was not true, confirming the higher resolution provided by pan-PCR for this test set (Fig. 4).

To provide a more realistic test set, we next evaluated pan-PCR and MLST on a set of six strains isolated during outbreaks at NIHCC in 2007 and 2009 (37). The 2007 outbreak was characterized by three PFGE strain types (A, B, and C), all of which belonged to the European clone II lineage. The 2009 outbreak was composed of a single PFGE strain of type D, which was a member of the European clone I lineage. We found that while both MLST and pan-PCR could differentiate types A, B, and C from type D, only pan-PCR could differentiate A, B, and C (Fig. 5A). Furthermore, pan-PCR supports the close phylogenetic relationship between strains A and non-multidrug resistant A (Anon-MDR) (Fig. 5B). Anon-MDR is a drug-sensitive strain present in NIHCC a year before the outbreak, which we previously postulated evolved into strain A (16). Even PFGE was not able to make this connection, indicating that pan-PCR provides sufficient resolution to distinguish independent strains from the same clonal lineage, while still being able to capture relationships among related strains.

**Comparison of assay resolution with that of optical mapping.** As a final assessment of our assay, we compared it with optical mapping-based typing. Optical mapping using the OpGen technology creates an ordered restriction map (see Fig. S4 in the supplemental material), which roughly equates to a digital PFGE map. We applied our PCR assay and OpGen to an additional set of 10 A. baumannii isolates, taken from patients in the NIH Clinical Center between 2009 and 2012. This set of isolates was selected to represent a more realistic application of this approach, whereby relationships among strains found in a hospital over a short period of time are determined. Application of pan-PCR to these 10 A. baumannii isolates categorized the genomes into a few distinct groups (see Fig. S5 in the supplemental material). Furthermore, the distances between isolates determined by the PCR assay and OpGen were highly correlated (Fig. 6), indicating that the PCR assay provides strain relationship information comparable to that provided by both optical mapping and PFGE.

**DISCUSSION**

We have designed a computational framework, pan-PCR, which exploits the exponentially increasing numbers of fully sequenced bacterial genomes to automatically design multiplex PCR typing assays based on species pan-genomes. To test this framework, we applied pan-PCR to Acinetobacter baumannii to generate a six-gene assay capable of distinguishing all input strains. Experimental validation confirmed that the computationally predicted presence/absence profiles could be replicated experimentally in a multiplex assay. Comparisons of pan-PCR’s discriminatory power with the discriminatory powers of MLST (4, 5) and optical mapping (27, 28) demonstrated that it is able to place distantly related strains in context, while providing the resolution necessary to distinguish closely related strains.

In general, PCR-based typing assays, such as those designed for Streptococcus pneumoniae (13), Neisseria meningitidis (14), and Acinetobacter baumannii (15), are desirable because of their inherent advantages of low cost and high speed. The expense comprises the cost of standard PCR reagents, and the turnaround time consists of only the few hours needed to prepare DNA and run the PCR. However, design of such assays by hand can be a cumbersome process, and the resulting assay may involve a large number of PCR targets, which can be unwieldy. Pan-PCR uses computational power to automate the design of multiplex PCR assays that require a near minimal number of cleanly spaced targets. To facilitate broad applicability, pan-PCR was designed to allow easy design of typing assays, without requiring bioinformatics expertise. Recommended parameters are suggested as default settings; however, users may adjust these parameters to adapt pan-PCR for their purposes. The pipeline to pick target regions and design primers from the input set of 29 A. baumannii genomes was completed in less than 30 min on a standard desktop computer, demonstrating that a sophisticated computational infrastructure is not required to run pan-PCR. Once an assay is designed, an initial reference database of pan-PCR profiles can be generated computationally from the sequenced genomes. This database can then be
supplemented by individual clinical labs with experimentally determined pan-PCR profiles of strains previously observed in the hospital.

While the approach was piloted on *A. baumannii*, pan-PCR can be used for any species with an available set of reference genomes. Comparative genomic studies of *Escherichia coli* (38, 39), *Salmonella* (40), and *Enterococcus faecalis* (41) indicate the presence of a great diversity in gene content that can be exploited by pan-PCR to design discriminatory assays. Furthermore, as the number of publicly available genomes increases, subspecies assays can be designed to provide additional discriminatory power. The flexibility of pan-PCR’s framework will seamlessly accommodate this application by altering only the set of input genomes. This will allow the design of assays targeting specific clonal lineages of interest. Moreover, as the cost of bacterial genome sequencing continues to decrease, it will become feasible for individual microbiology laboratories to sequence a set of genomes to augment the pool of sequenced strains of interest.

Finally, an issue of importance with any typing scheme is the ability to define a discriminatory threshold by which two strains are declared distinct. In the context of infection control, having a robust discriminatory threshold is critical in attempting to discern whether a patient isolate is the product of a new introduction or a nosocomial transmission. As bacterial genomes are constantly evolving and, hence, bacterial types are continually changing, what makes for a robust assay is one in which bacterial types change in a steady and predictable manner. We believe that pan-PCR, because of its basic design principles, imparts some degree of stability in the face of evolution, as it filters out genomically unstable elements such as prophages, transposons, and hypothetical proteins. In addition, because the gene selection process inherently favors the selection of genes that move in and/or out of genomes independently of one another, multiple gain/loss events should be required for multiple changes in banding patterns. In contrast, it is unclear how drastically a PFGE banding pattern can change from a single genomic rearrangement. MLST may also be subject to rapid changes in sequence types, as recent work has shown that some organism’s propensity for large recombination events may complicate MLST comparisons (16).

The sequencing revolution has now entered a phase where the power of genomics is beginning to be harnessed for clinical applications. We provide a framework that allows microbiology laboratories of all sizes to exploit genome sequence data to design high-resolution, cost-effective, and rapid-result bacterium-typing assays. Empowering microbiology laboratories to capitalize on the genome revolution will contribute to more effective treatment and containment of bacterial infections.

**ACKNOWLEDGMENTS**

Karen Frank provided thoughtful comments and discussion. We thank the NIH Clinical Center Department of Laboratory Microbiology for technical assistance.

Research support came from NHGRI and NIHCC intramural research programs. E.S.S. is supported by a Pharmacology Research Associate Training Fellowship, NIGMS.

**REFERENCES**