A Phylogeny-Informed Proteomics Approach for Species Identification within the *Burkholderia cepacia* Complex

Honghui Wang, Ousmane H. Cissé, Thomas Bolig, Steven K. Drake, Yong Chen, Jeffrey R. Strich, Jung-Ho Youn, Uchenna Okoro, Avi Z. Rosenberg, Junfeng Sun, John J. LiPuma, Anthony F. Suffredini, John P. Dekker

*Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA*
*Proteomics Core Facility, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA*
*Department of Laboratory Medicine, Clinical Center, Microbiology Service, National Institutes of Health, Bethesda, Maryland, USA*
*Kidney Disease Section, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, USA*
*Department of Pathology, Johns Hopkins University, Baltimore, Maryland, USA*
*Department of Pediatrics, University of Michigan Medical School, Ann Arbor, Michigan, USA*
*Bacterial Pathogenesis and Antimicrobial Resistance Unit, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA*

Authors Honghui Wang and Ousmane H. Cissé contributed equally to this work. Co-first author order was determined by relative contributions to the clinical assay design; H.W. designed the clinical assay and O.H.C. performed the supporting genomics analysis.

ABSTRACT Ancestral genetic exchange between members of many important bacterial pathogen groups has resulted in phylogenetic relationships better described as networks than as bifurcating trees. In certain cases, these reticulated phylogenies have resulted in phenotypic and molecular overlap that challenges the construction of practical approaches for species identification in the clinical microbiology laboratory. *Burkholderia cepacia* complex (Bcc), a betaproteobacteria species group responsible for significant morbidity in persons with cystic fibrosis and chronic granulomatous disease, represents one such group where network-structured phylogeny has hampered the development of diagnostic methods for species-level discrimination. Here, we present a phylogeny-informed proteomics approach to facilitate diagnostic classification of pathogen groups with reticulated phylogenies, using Bcc as an example. Starting with a set of more than 800 Bcc and *Burkholderia gladioli* whole-genome assemblies, we constructed phylogenies with explicit representation of inferred interspecies recombination. Sixteen highly discriminatory peptides were chosen to distinguish *B. cepacia*, *Burkholderia cenocepacia*, *Burkholderia multivorans*, and *B. gladioli* and multiplexed into a single, rapid liquid chromatography-tandem mass spectrometry multiple reaction monitoring (LC-MS/MS MRM) assay. Testing of a blinded set of isolates containing these four *Burkholderia* species demonstrated 50/50 correct automatic negative calls (100% accuracy with a 95% confidence interval [CI] of 92.9 to 100%), and 70/70 correct automatic species-level positive identifications (100% accuracy with 95% CI 94.9 to 100%) after accounting for a single initial incorrect identification due to a preanalytic error, correctly identified on retesting. The approach to analysis described here is applicable to other pathogen groups for which development of diagnostic classification methods is complicated by interspecies recombination.

KEYWORDS clinical microbiology, computational biology, genomics, mass spectrometry, network phylogeny, proteomics

The genus *Burkholderia* is a betaproteobacteria group comprising more than 120 species with remarkably diverse lifestyles, ranging from plant pathogens to significant human pathogens, such as *Burkholderia pseudomallei* (1, 2). Nested within the
Burkholderia genus, the Burkholderia cepacia complex (Bcc) is a highly polymorphic clade whose members cause life-threatening chronic respiratory infections in persons with cystic fibrosis, along with opportunistic infections in certain immunocompromised hosts, such as those with chronic granulomatous disease (3–6). Diagnostic classification of species within the complex is epidemiologically and clinically important due to varying pathogenicity and antibiotic resistance profiles (7, 8). For example, the identification of Burkholderia cenocepacia infection in cystic fibrosis lung transplant recipients is associated with worse posttransplant survival compared to other Bcc species (9, 10). Furthermore, Burkholderia dolosa, B. cenocepacia, Burkholderia multivorans, and B. cepacia have higher rates of antibiotic resistance, and early identification of these pathogens may help guide antibiotic therapy (8).

Despite the clinical importance of this pathogen group, accurate diagnostic classification in the hospital microbiology laboratory has proven challenging due to the combination of phenotypic overlap between species and a high level of intraspecies diversity, which are, in part, consequences of extensive genetic exchange between members of the Bcc (11–13). Previous work has suggested that this history of genetic exchange in the core genome has resulted in a phylogeny with significant network structure (14, 15). Other work has documented translocation and parallel rearrangement events both within and between the replicons (chromosomes) of the Bcc core genome, with complex patterns of positive and negative selection (16). This complexity results in an important additional challenge in identifying and confirming discriminatory diagnostic targets in genomic data sets that may undersample the true diversity (17). Despite this underlying complexity, phylogenies have been constructed based on small numbers of marker genes, including single marker gene approaches such as recA gene-based phylogenies, and multiple marker gene approaches, such as multilocus sequence typing (MLST)-based phylogenies (18–21). A widely used scheme is based on seven markers, typically including recA, gyrB, and rpoB that can be used in conjunction with 16S rRNA sequences (22). Such schemata, however, can provide inconsistent species determinations, with consequences for diagnostic classification of clinical isolates (23).

Mass spectrometry (MS) has come to play an important role in clinical microbiology laboratories, with matrix-assisted laser desorption–ionization time of flight (MALDI-TOF) MS now used routinely in hospital laboratories for rapid identification to species of bacterial pathogens. However, commercial MALDI-TOF MS approaches, in which identification is based on conserved ribosomal and heat shock proteins, have difficulty differentiating species within the Bcc, similar to marker gene-based methods (13, 24). Previously we developed a genoproteomic approach using liquid chromatography-tandem mass spectrometry (LC-MS/MS) to identify Gram-negative bacteria at the species and strain levels (25, 26). This method combines theoretical peptidome analysis from in silico translation and tryptic digestion of bacterial whole-genome sequences followed by experimental selection of optimal peptides for identification by LC-MS/MS. Similar approaches have been used for the detection of antibiotic resistance proteins in bacteria (27–29).

In this work, we present a general strategy for the development of rapid mass spectrometry diagnostic assays to classify clinical pathogens with reticulated phylogenies, using Bcc as an example. The approach combines phylogenetic reconstruction with experimental peptidomic analysis, and the resulting assays can be implemented on mass spectrometry instruments commonly available in clinical laboratories. Starting with a large set of Bcc and Burkholderia gladioli whole-genome assemblies, we constructed phylogenetic trees with inferred interspecies recombination explicitly represented. We selected 16 highly discriminatory peptides and used these peptides to build an automated multiplex LC-MS/MS MRM assay. The phylogenetic and proteomic methods developed here would be applicable to other pathogen groups exhibiting reticulated phylogenies, such as Campylobacter, Helicobacter, Neisseria, and others (30–35).
MATERIALS AND METHODS

Large-scale Burkholderia genome analysis. We analyzed publicly available whole-genome sequences of 893 Burkholderia species downloaded from the NCBI database. NCBI classification of these species is presented in Table S1 in the supplemental material. The vast majority of the genomes were sequenced using Illumina technology. Since only a minority had available annotations, we annotated all genome assemblies and predicted ~7,000 protein-coding genes per genome assembly. Given that the average Burkholderia genome size is 7.6 Mb, genomes with sizes <7 Mb were considered potentially incomplete (n = 2). Potentially nonpublic genomes were excluded to the best of our ability. No sampling strategy was used in the study.

Genome annotation, gene family identification, and phylogeny. Genome sequences were annotated using GenMarkS version 4.32 and screened for the presence of stop codons and frameshifts (36). Computing pairwise orthologs for >800 proteomes (each containing ~7,000 coding DNA sequences [CDS]) was computationally prohibitive. Instead, 100 genomes were selected randomly and orthology was established using reciprocal best BLASTp hit (E value 10^-10) (37). The clustering yielded 14,926 groups. For each group, we extracted the best homolog from the remaining predicted proteomes using BLASTp (E value of 10^-10) and a coverage threshold of >80%. Duplicate entries (potential paralogs) were collapsed. Multiple sequence alignments were generated using MUSCLE and subsequently filtered with TrimAl using the automated option (38, 39). Each alignment was screened with PhiPack for intragenic recombination (40). Transposons were identified with transposonPSI (http://transposonpsi.sourceforge.net) and by BLAST against a custom database built from UniProt (41). Approximate maximum likelihood phylogenetic trees were estimated using FastTree with the GTR+CAT (general time reversible with per site rate categories) model of approximation for site rate variation with 1,000 resampling times (42). We extracted seven canonical markers used for MLST typing, and maximum likelihood (ML) trees were generated using IQ-TREE (-m TEST -b 100 -nt 16 -a S 1000) (43). In parallel, we concatenated the alignments into a single super alignment and an ML tree was inferred using IQ-TREE.

A species tree was inferred using ASTRAL-II (44). Tree comparisons were performed using ETE3 and/or Phangorn (45–47).

Analysis of recomposition. We generated a superalignment by merging prealigned orthologous groups containing a single best homolog from each assembly. Only complete groups were retained. After filtering, 652 assemblies, each containing 880 homologs, were included in the superalignment. An approximate likelihood tree for the superalignment was generated with FastTree. ClonalFrameML (version 1.11.3) software was run with 100 simulations to estimate uncertainty in the Baum-Welch approximate likelihood tree for the superalignment was generated with FastTree. ClonalFrameML

Bacterial isolates. All bacterial isolates used in this study are listed in Table S2. Thirty-three isolates used in assay development were deidentified clinical isolates obtained from the NIH Clinical Center collection and/or from the ATCC. Initial proteomic work was based on ATCC 25416 (B. cepacia), ATCC BAA-245 (B. cenocepacia), ATCC BAA-247 (B. multivorans), and a deidentified B. gladioli clinical isolate. The remaining isolates were Burkholderia stabilis, Burkholderia stabilis, and Burkholderia vietnamiensis isolates were used as other negative-control Burkholderia species for assay development. Eighty of the isolates used for blinded evaluation were obtained from the B. cenocepacia Research Laboratory and Repository at the University of Michigan Medical School.

Protein extraction and tryptic digestion. Bacterial isolates were grown on blood agar plates (Remel, Lenexa, KS) for 18 to 24 h at 35°C with 5% CO2. Proteins were prepared as previously described (51). Briefly, for each sample, a 10-μl loop of fresh bacterial cells was resuspended in 0.5 ml 70% ethanol, vortexed for 1 min, and centrifuged at 20,800 × g for 2 min. Supernatant was removed and the pellet was resuspended in 100 μl of 70% formic acid (FA) and mixed to homogeneity, followed by addition of 100 μl of 100% acetonitrile (ACN). The resulting solution was re-vortexed for 10 s and centrifuged for 2 min at 20,800 × g. Aliquots of supernatant (FA/ACN lysate) were stored at −20°C for later use. A mixture of 2 μl of protein lysate and 8 μl H2O in a 1.5-ml tube (Protein LoBind Tube, Eppendorf) was frozen in dry ice, then lyophilized for 20 min. The lyophilized protein was resuspended in 96 μl of 50 mM NH4HCO3 and sonicated for 2 min (Qsonica Q500) with a pulse of 20 s on followed by 20 s off at an amplitude of 50 W. Then, 4 μl of 0.1 μg/μl of Trypsin Gold (Promega, Madison, WI) was added and digestion was carried out for 15 min at 55°C in a water bath. After digestion, the solution was spun down in an Ultrafree-MC GV centrifugal filter (Merck Millipore Ltd., Tullagreen, Ireland) for 3 min at 20,800 × g. Ten microliters of the digests were used for total protein concentration measurement by Qubit 2.0 and protein concentrations of 100 μg/ml were adjusted to 100 μg/ml with 50 mM NH4HCO3. For samples analyzed on the Agilent 6460 ESI LC-MS, 20 μl of lysate was processed with a speedvac for 6 min to eliminate ACN and formic acid, and then 20 μl of phosphate-buffered saline (PBS) was mixed with the remaining sample. This was frozen in dry ice and processed similarly to above except with 4 μg trypsin (1 μg/μl) used for the digests.

Lowest common ancestor analysis. For initial preliminary investigation, the Unipept web tool pept2lca was used to perform lowest common ancestor (LCA) analysis for every peptide of the core
peptidomes (defined as the set of peptides present in >85% of genomes in each species) of \textit{B. cepacia}, \textit{B. cenocepacia}, \textit{B. multivorans}, and \textit{B. gladioli}. Peptides were defined as species-specific if they were found in the genome of only one species by LCA analysis. The set of identified core peptides that were found by LCA to be species-specific for \textit{B. multivorans}, \textit{B. gladioli}, \textit{B. cepacia}, and \textit{B. cenocepacia} were then selected as potential unique peptides for sensitivity and specificity calculations as outlined below. In addition, since the number of species-specific peptides in \textit{B. cepacia} and \textit{B. cenocepacia} was low, we also considered the set of Bcc-specific peptides in the proteomes of these two species. We reasoned that this set of Bcc-specific peptides might contain peptides that were, in fact, specific to \textit{B. cepacia} or \textit{B. cenocepacia}, but that had been misclassified by LCA as shared peptides due to misclassifications between \textit{B. cepacia} and \textit{B. cenocepacia} in the underlying database.

**Estimation of diagnostic sensitivity and specificity for selected species-specific peptides.** Sensitivity and specificity of selected species-specific and Bcc-specific peptides for discrimination among \textit{B. cepacia} and \textit{B. cenocepacia} were estimated as follows. All of the 859 core peptides that were assigned as Bcc-specific within the \textit{B. cepacia} proteome were searched against 874 \textit{Burkholderia} genomes to determine (i) how many of the 87 \textit{B. cepacia} genomes (based on our phylogenetic network classification) contain this peptide (sensitivity) and (ii) how many of 787 genomes that were not classified as \textit{B. cepacia} lack this peptide (specificity). The same approach was applied to find the sensitivity and specificity for each of the 1,122 peptides that were assigned as Bcc-specific within the \textit{B. cenocepacia} proteome. Based on this classification, the sensitivity and specificity for each core peptide for species-specific discrimination was calculated among 874 genomes. For final selected core peptides for the MRM assay, the predicted sensitivity of each selected peptide was >95% and predicted specificity was >99%.

**Experimental identification of peptides.** The tryptic-digested solutions of the 4 species of interest (\textit{B. cenocepacia}, \textit{B. cepacia}, \textit{B. multivorans}, and \textit{B. gladioli}) were diluted 160-fold and then 2 \mu l and 10 \mu l of the diluted solutions were injected to a Thermo Orbitrap Fusion LC-MS/MS to identify unique peptides. Peptides that were detected in both injections were considered optimal candidates (i.e., highly abundant, efficiently ionized, and robustly detected). LC-MS/MS data were searched against a custom database composed of \textit{B. cepacia} (ATCC 25416), \textit{B. cenocepacia} (ATCC BAA-245), \textit{B. multivorans} (ATCC BAA-247), and \textit{B. gladioli} (ATCC 10248) FASTA files by Proteome Discoverer 2.2 (Thermo Fisher Scientific, Waltham, MA). The FASTA files were downloaded from NCBI. Label-free quantification was analyzed based on the peak intensity of the precursor ion.

**Screening of the potential peptide markers.** Table S2 lists isolates that were included in the MRM assay development set. Potential peptide markers were screened by MRM without labeled peptides using tryptic-digested samples from the following reference isolates: ATCC 25416 (\textit{B. cepacia}), ATCC BAA-245 (\textit{B. cenocepacia}), ATCC BAA-247 (\textit{B. multivorans}), and NIH1 (\textit{B. gladioli}). The peptides were selected based on detected MRM spectral intensity, retention time, peak shape, lack of carryover, and interferences.

**Labeled peptides.** Peptide standards containing heavy isotopic labels in R (U-13C6; U-15N4) or K (U-13C6; U-15N2) C-terminus amino acids were purchased from JPT (Berlin, Germany) with characterization and concentration provided by the manufacturer. The labeled peptides were stored in 50 mM or 100 mM NH4HCO3 at 10 pm/\mu l or 1 pm/\mu l at −20°C and diluted with 50 mM or 100 mM NH4HCO3 as required.

**Quality control sample.** A mixture of equal volumes of the tryptic digests from four samples representing \textit{B. cepacia}, \textit{B. cenocepacia}, \textit{B. multivorans}, and \textit{B. gladioli} was prepared and stored at −20°C until use in 30-\mu l aliquots for the purpose of quality control (QC) samples. The QC sample was run before sample 1 and after samples 25, 50, 75, 100, and 120 to ensure that all 16 peptides were detected and their retention times were within the retention window set in the MRM assay.

**MRM method optimization.** A rapid, dynamic multiple reaction monitoring (MRM) LC-MS/MS assay was developed and optimized for these peptides on an Agilent ChipCube 6495 Triple Quadrupole LC/MS system with a high-capacity chip. Mobile phases A and B were 0.1% FA, 5% ACN, 95% H2O, and 0.1% FA, 95% ACN, 5% H2O, respectively. The gradient was run from 5% B to 30% B in 10 min with a 0.4 \mu l/min flow rate. The total assay time was 20 min. The retention time window was set at 2 min with the Q1 and Q3 mass resolution settings shown in Table 1. Other settings for MS included the following: Delta EMV+ = 300; cell accelerator voltage: 3: gas temperature 200°C, and gas rate of 11 liter/min. Table 1 lists the peptides and transitions as well as collision energy for each transition. The labeled peptide concentrations (Table 1) were calculated based on the manufacturer’s specifications. Labeled peptide mix (3 \mu l) was added to 30 \mu l of digested peptide solution in a silanized vial (National C4000-S9, Thermo) or a high-recovery vial (Chemglass Life Science, CV-5193-1232). An aliquot of 4 \mu l was injected into the LC-MS. Between the sample runs, a no-matrix “blank” with 0.2 \mu l of a labeled peptide mixture was injected to the column as a quality control to monitor the LC-MS performance during batched runs. The column was washed and re-equilibrated with a 30-min washing protocol at the end of the day.

**Analytical detection thresholds.** To establish detection thresholds of the assay during test development, tryptic digests of a \textit{B. cepacia} isolate were mixed with that of \textit{B. cenocepacia} or vice versa in a ratio that ranged from 1 to 128, and the same experiment was performed for \textit{B. multivorans} and \textit{B. gladioli}. The detection threshold varied among the peptides, but at least one peptide per species, including LDDDLLAVADGNR (\textit{B. cepacia}), TLGVAPGTGFGEF (\textit{B. cenocepacia}), DLQADQDDDLR (\textit{B. multivorans}), and ASEFAYAVLNDVK (\textit{B. gladioli}) was still clearly detectable following a 64-fold dilution. Figure S1 shows representative LC-MS chromatograms for these four peptides after 64-fold dilution. Peptides GLGTOKPTPLDVNVGTGR (\textit{B. cepacia}) and ESIVDGAALPIK (\textit{B. cenocepacia}), in contrast, demonstrated higher detection thresholds with more interferences.

**Blinded MRM assay evaluation.** An Agilent 6460 Triple Quadrupole LC-MS/MS system with 1290 infinity II ultra-high-performance liquid chromatography (UHPLC) was used for the blinded evaluation of
**TABLE 1** Peptides and transitions used in the MRM assay

<table>
<thead>
<tr>
<th>Species</th>
<th>Peptide sequence</th>
<th>Precursor ion</th>
<th>RT (min)</th>
<th>T1&lt;sup&gt;b&lt;/sup&gt;</th>
<th>T2&lt;sup&gt;b&lt;/sup&gt;</th>
<th>T3&lt;sup&gt;b&lt;/sup&gt;</th>
<th>T4&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. cenoceapa</td>
<td>ESIVDGAALPIK</td>
<td>606.8454</td>
<td>2.5</td>
<td>883.5247 (22)</td>
<td>669.4294 (23)</td>
<td>357.2496 (19)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPAGHPAQLPLAAGTAAR</td>
<td>533.6302</td>
<td>1.2</td>
<td>772.4100(17)</td>
<td>617.3365 (16)</td>
<td>546.2994 (18)</td>
<td>386.7087 (16)</td>
</tr>
<tr>
<td></td>
<td>LTPITLPMTAPFGK</td>
<td>808.4418</td>
<td>3.2</td>
<td>1062.5652 (30.1)</td>
<td>848.4335 (29.1)</td>
<td>701.3760 (25.1)</td>
<td>448.2554 (29.1)</td>
</tr>
<tr>
<td></td>
<td>TGLVGAPGTFGEFAK</td>
<td>747.4036</td>
<td></td>
<td>1123.5782 (27.2)</td>
<td>1024.5098 (22.2)</td>
<td>953.4727 (26.2)</td>
<td>371.2289 (23.2)</td>
</tr>
<tr>
<td>B. cecapica</td>
<td>GLITKPPTLVDNVWVTGR</td>
<td>632.0301</td>
<td>2.6</td>
<td>862.4887 (17)</td>
<td>777.4395 (19)</td>
<td>603.3209 (22)</td>
<td>390.2096 (14)</td>
</tr>
<tr>
<td></td>
<td>LTTDDOLAVAGDNR</td>
<td>687.8284</td>
<td>1.5</td>
<td>702.3529 (25.3)</td>
<td>631.3158 (24.3)</td>
<td>532.2474 (25.3)</td>
<td>346.1833 (26.3)</td>
</tr>
<tr>
<td></td>
<td>ADDAPEGQAPAEEDGPR</td>
<td>646.6243</td>
<td>0.8</td>
<td>783.3688 (20.5)</td>
<td>606.2918 (22.5)</td>
<td>373.1334 (17.5)</td>
<td>302.0983 (21.5)</td>
</tr>
<tr>
<td></td>
<td>AAVDVLATQAPPR</td>
<td>661.8750</td>
<td>2.1</td>
<td>853.4890 (25.5)</td>
<td>740.4049 (23.5)</td>
<td>568.3202 (19.5)</td>
<td>440.2616 (17.5)</td>
</tr>
<tr>
<td></td>
<td>AISTIQLDTGAAPGSSPG</td>
<td>662.0117</td>
<td>1.7</td>
<td>900.4534 (23)</td>
<td>749.8715 (23)</td>
<td>685.8422 (23)</td>
<td>629.3002 (23)</td>
</tr>
<tr>
<td>B. multivorans</td>
<td>DLQADVQQDDL</td>
<td>644.3124</td>
<td></td>
<td>931.44795 (25)</td>
<td>860.4108 (23)</td>
<td>745.3839 (25)</td>
<td>646.3155 (22)</td>
</tr>
<tr>
<td></td>
<td>FVLETADDQADPR</td>
<td>738.8519</td>
<td>1.9</td>
<td>1230.5597 (26.9)</td>
<td>1117.4756 (24.9)</td>
<td>988.4330 (24.9)</td>
<td>887.3853 (26.9)</td>
</tr>
<tr>
<td></td>
<td>NLNLNFPLDSMK</td>
<td>746.8769</td>
<td>3.1</td>
<td>1265.6194 (28.2)</td>
<td>1038.4925 (25.2)</td>
<td>924.4495 (27.2)</td>
<td>777.3811 (27.2)</td>
</tr>
<tr>
<td></td>
<td>TGLVAPGTFGDFA</td>
<td>754.3988</td>
<td>3</td>
<td>1137.5687 (27.4)</td>
<td>1038.5003 (28.4)</td>
<td>967.4632 (28.4)</td>
<td>272.1605 (27)</td>
</tr>
<tr>
<td>B. gladiolii</td>
<td>ASEYAAVLNDVK</td>
<td>640.3301</td>
<td>2.2</td>
<td>992.5411 (23)</td>
<td>829.4778 (25)</td>
<td>758.4407 (21.9)</td>
<td>588.4407 (19)</td>
</tr>
<tr>
<td></td>
<td>AVELAVEQLPVPVER</td>
<td>569.9966</td>
<td>2.5</td>
<td>769.4385 (16)</td>
<td>704.9172 (19)</td>
<td>648.3751 (18)</td>
<td>612.8566 (18)</td>
</tr>
<tr>
<td></td>
<td>LTSEAIQSFR</td>
<td>576.3064</td>
<td>1.8</td>
<td>937.4738 (19.9)</td>
<td>850.4417 (22.9)</td>
<td>721.3991 (21.9)</td>
<td>650.3620 (19.9)</td>
</tr>
</tbody>
</table>

<sup>a</sup>RT, retention time using UHPLC and Agilent 6460 ESI-LC-MS/MS QQQ.

<sup>b</sup>T1 to T4, transition masses in m/z. Collision energies are given in parentheses.

the assay. Mobile phases A and B were 0.1% FA, 2% ACN, 98% H₂O, and 0.1% FA, 100% ACN, respectively. The gradient was run from 10% to 30% B in 3 min with 0.5 ml/min flow rate. The total assay time was 6 min. The MRM method was exactly same as that developed for use on the ChipCube QQQ, except for the retention time and retention time windows. Retention times for each peptide appropriate to the Agilent 6460 Triple Quadrupole LC-MS/MS were determined using the labeled peptide mix and a retention window was set to 0.4 min for all transitions. The column used was HALO 2 Peptide ES-C18 2.1 × 100 mm 2 μm. During the blinded evaluation, the QC sample was run before sample 1 and after samples 25, 50, 75, 100, and 120.

To prepare samples for the blinded evaluation, 6 μl of labeled peptide mix was added to 40 μl of 1 μg/μl peptide digest (to make 46 μl total), and 20 μl of this mixture was injected into the LC-MS. This reflected a 50× increase in sample and labeled peptide relative to what was used during test development on the Agilent 6495 ChipCube QQQ to compensate for the sensitivity loss with electrospray ionization (ESI) compared to nanospray ionization (NSI). After every 10-sample set, a blank was run to clean the column using the same gradient with an additional 4 min of 70% B. For the blinded MRM assay evaluation, acquired data were analyzed with Skyline 19.1 software (MacCoss Lab), and the transition window boundaries for each peptide in each sample were verified and adjusted as necessary by a single blinded operator (Operator 1). Operator 1 executed a script based on the automatic individual peptide call rules to make positive and negative individual peptide calls. Individual peptides (total 10 peptides) marked for manual inspection were examined by three operators (operators 1 to 3), including examination of LC-MS chromatograms in Skyline, and manual calls as positive or negative were made for each of the peptides. For all peptides that had been called either by automatic or manual methods, operator 2 used a second Excel script to convert the Excel file into a table that could be used to make final species-level calls for all specimens with at least two individual species-specific peptides that were called positive.

**Statistical analysis.** Exact 95% confidence intervals throughout were calculated using the Clopper-Pearson method implemented in R package `binom`.

**Data availability.** All genomic data are available at: https://doi.org/10.5281/zenodo.3758670 and https://github.com/ocisse/Burkholderia_recombination.

**RESULTS**

**Analysis of Burkholderia genomes reveals frequent recombination and massive gene flow.** Recombination between species can result in complex network architecture and distort phylogenetic trees, as has been previously described (52). To gain a better understanding of the impact of recombination within the *Burkholderia* genus, we downloaded 893 *Burkholderia* genomes from the National Center for Biotechnology Information (NCBI) (Table S1) representing isolates from Australia (*n* = 445), Canada (*n* = 198), USA (*n* = 59), Thailand (*n* = 56), and other/unknown locations (*n* = 135). After exclusion of incomplete genomes and annotation of protein-coding regions, we extracted 2,545 orthologous groups, including 2,403 intragenic recombination-free markers and 142 recombiant markers as inferred from a pairwise homology index (PHI) with 10,000 permutations using PhiPack (40). We then generated coalescent
multispecies trees with and without the 142 recombinant markers using ASTRAL-II, which allows the use of partial gene trees and is robust to incomplete lineage sorting and confused gene tree signals (44, 53). By comparing trees built with and without intragenic recombinants, we found that recombination does significantly impact the topology of the trees, with these trees differing by 14% of their maximum distance (Robinson-Foulds distance of 0.144).

We next explored the effects of recombination at a more detailed level using ClonalFrameML, a maximum likelihood method applicable to large genomic data sets, and BACTER, a Bayesian method suitable only for small data sets (48, 54). To facilitate analysis with ClonalFrameML, we generated a superalignment of 1.9×10^9 nucleotides by concatenating filtered prealigned orthologous groups derived from 652 assemblies (880 genes per assembly). We then generated a maximum likelihood tree from the superalignment using FastTree to use as a seed for ClonalFrameML (42). The impact of recombination on Burkholderia genomes was quantified and the mean parameters were estimated to be R/theta = 0.144; mean DNA import length per branch (l) = 187.8 bp; and ν = 7.73×10^{-2}. These data suggest massive gene flow among the genomes used in this analysis (Fig. 1, top). We also noted variation in the amount of gene flow between different branches of the tree, which may suggest that some branches are more prone to recombination than others. Ancestral recombination graphs (ARG) generated by BACTER analysis of six whole genomes, including B. cepacia, B. cenocepacia, B. multivorans, B. vietnamiensis, B. stabilis, and B. gladioli, and an analysis of seven legacy MLST markers (atpD, gltB, gyrB, lepA, phaC, recA, and trpB) also reveal significant conversion events in our data set (Fig. 1, bottom). A summary ARG using the seven MLST markers with a conversion posterior cutoff of 0.6 found that many conversion events have posterior support that exceeds this cutoff.

Network constructions capture an extensively reticulated Burkholderia phylogeny. To construct a robust phylogenetic classification, we built a network based on a neighbor-joining algorithm that takes a distance matrix as input and infers relationships by agglomerating clusters (55). Computing a network, including 652 assemblies (880 genes per assembly), was computationally intractable in our hands. We attempted unsuccessfully to generate a phylogenetic network of 880 genomes using alternative methods, including a consensus network based on complete gene trees and a method based on partial trees (56, 57). Failures were caused by the inability to handle partial gene trees for the former and the overrepresentation of splits for the latter, which makes the computation intractable.

To reduce the computational burden, we divided our data set into blocks of ~200 assemblies/locus (where each locus is a concatenation of 880 genes). These blocks were assembled to include clades of interest that were selected based on pairwise distances with reference genomes (Fig. S2). Each block was analyzed separately to produce a phylogenetic network, and then combined. Figure 2 and Fig. S3 demonstrate the resulting individual networks and highlight the substantially reticulated phylogenetic structure that is present.

Network phylogeny facilitates species groupings for the selection of diagnostic peptides. The phylogenetic analysis performed above demonstrated substantial recombination between the analyzed Bcc species. To facilitate selection of highly discriminatory peptides for diagnostic identification, we used a manual curation approach in which qualitative boundaries were introduced to separate most of the core reticulations from the species-specific branches in the individual networks (Fig. 2). Placement of these boundaries was based on expert judgement and guided by an assumption that NCBI genome classifications contained some number of misidentifications (which the network was intended to correct), but that the majority of NCBI classifications were accurate. As a result of boundary placement, 87 genomes were grouped with B. cepacia, 244 genomes were grouped with B. cenocepacia, 29 genomes were grouped with B. multivorans, and 10 genomes were grouped with B. gladioli. We then compared the species groupings provided by this analysis with the species assignments for each genome given in NCBI, which revealed a few important differences. Four genomes
grouped with *B. cepacia* in the analysis above were annotated in the NCBI database as non-*B. cepacia* (LDWF [*B. lata*], LDWP01 [*B. lata*], ASSI01 [*Burkholderia* sp., AU4i], and LDUL01 [*Burkholderia* sp., LK4i]), and three genomes grouped with *B. cenocepacia* by the analysis above were annotated in the NCBI database as *B. cepacia* (GCA_000755805, JPE01, and JTD01). Finally, one genome classified by the
network analysis as *B. gladioli* (JJMR01) was annotated as *Burkholderia* sp., A1 in the NCBI database. All 29 genomes that grouped with *B. multivorans* were classified as *B. multivorans* in NCBI.

**Prediction of theoretical peptides and experimental identification with Orbitrap LC-MS/MS.** Given the differences above in classification of *B. cepacia* and *B. cenocepacia* genomes between NCBI and the network phylogeny, we based our analysis on the classifications provided by the latter. With these classifications, we generated lists of core peptides for each species, ranging from 30,000 to 72,000, and defined as the set of peptides present in greater than 85% of the genomes for each group. In the next step, we used an Orbitrap Lumos LC-MS/MS instrument in routine data dependent acquisition (DDA) mode to detect proteins and peptides in four selected isolates representing *B. cepacia*, *B. cenocepacia*, *B. multivorans*, and *B. gladioli* (Table S3 and Table S4). Peptides detected with high confidence using the default definition on Proteome Discoverer 2.2 were filtered to include only the core tryptic peptides identified above that had a peptide length of 8 to 25 moieties with no missed cleavages or oxidation. A total of 6,888, 6,306, 2,446, and 4,797 core peptides were detected for *B. cepacia*, *B. cenocepacia*, *B. multivorans*, and *B. gladioli*, respectively.

**Identification of species-specific discriminatory core peptides.** We next sought to choose marker peptides with high discriminatory capacity for our assay from the set of identified core peptides. As an initial step in this analysis, we employed a simple lowest common ancestor (LCA) Unipept ([https://unipept.ugent.be/](https://unipept.ugent.be/)) metaproteomics approach. In this analysis, we looked for core peptides that were unique to each of the four *Burkholderia* species individually (“species-specific core peptides”) and those that were restricted to the Bcc as a group (“Bcc-specific peptides”). Note that the latter group excludes *B. gladioli*, which is not a member of the Bcc. This approach identified 170 and 1,271 core peptides as species-specific for *B. multivorans* and *B. gladioli*, respectively (Table S4). However, only 11 and 4 core peptides were found to be species-specific for *B. cepacia* and *B. cenocepacia*, respectively, an expected consequence of misclassifications of *B. cepacia* and *B. cenocepacia* genomes in the underlying database. We thus used species classifications from our phylogenetic network to assist in selection of discriminatory peptides. To do this, we took all core peptides that were assigned as Bcc-specific by LCA within the *B. cepacia* and *B. cenocepacia* proteomes and calculated individual peptide-level sensitivity and specificity parameters for species-level discrimination based on species grouping from the phylogenetic network. These parameters were calculated for 859 predicted Bcc-specific peptides within the *B. cepacia* proteome and 1,122 predicted Bcc-specific peptides within the *B. cenocepacia* proteome.
proteome. For *B. cepacia*, 265 out of the 859 predicted peptides were present in no more than 3 out of 787 non-*B. cepacia* genomes, indicating these 265 peptides were highly specific to *B. cepacia*. Furthermore, 89 out of these 265 highly specific peptides were present in at least 84 out of the 87 *B. cepacia* genomes, indicating these 89 peptides also had high sensitivity for *B. cepacia*. For *B. multivorans* and *B. gladioli*, the calculated specificities of the species-specific peptides matched well with the LCA analysis, reflecting a lesser degree of network structure interference in the phylogenetic separation of these groups.

**MRM LC-MS/MS assay development.** Based on the combination of the theoretical and experimental analysis above, we selected 16 peptides for the development of an MRM LC-MS/MS assay. Empiric factors considered included precursor ion peak intensities of >19,000K, MRM chromatograms, peak shapes, retention times, carryover, and the presence of interference. Table 2 lists the calculated sensitivity and specificity parameters for species-specific discrimination for each of the 16 selected peptides and Fig. 3 shows representative LC-MS chromatographs of all 16 peptides. Table 1 lists the peptides and transitions, as well as collision energy for each transition.

The peptide markers for *B. multivorans* and *B. gladioli* were predicted to have 100% sensitivity and specificity for species-level discrimination based on the above phylogenetic analysis. However, almost all individual peptide markers (except GLGITKPTLVVDNVGTR for *B. cepacia* and *B. cenocepacia*) had a sensitivity of less than 100%, though still greater than 95%. We thus calculated the sensitivity with two-peptide combinations (Table S5). These two-peptide combinations were found to increase the predicted coverage of *B. cepacia* and *B. cenocepacia* up to 100% for a number of combinations.

All peptides selected for the MRM assay were searched against the nonredundant NCBI database by protein BLAST (default parameters) to confirm that these peptides were not present in non-*Burkholderia* species (37).

To study the effects of sample dilution on detection thresholds of the assay, mixtures of *B. cepacia* and *B. cenocepacia* tryptic digests and of *B. multivorans* and *B. gladioli* were prepared in ratios that ranged from 1 to 128 (Fig. S1). Though the detection thresholds varied among the peptides, at least one peptide for each species, including LTDDDLAVADGNR (*B. cepacia*), TLGVAPGTFGEFAK (*B. cenocepacia*), DLQADVQDDLR (*B. multivorans*), and ASEYAAVLNDVK (*B. gladioli*), was clearly detectable following 64-fold dilution.

**Establishment and testing of automatic call rules for LC-MS/MS MRM assay.** To evaluate analytical performance and establish thresholds for automatic call rules, 33
FIG 3 LC-MS chromatograms for 16 discriminatory peptides used in the MRM assay. LC-MS chromatograms for 16 native (top) and labeled (bottom) peptides acquired by the MRM assay using UHPLC ESI-LC-MS, based on a sample consisting of equal volumes of the tryptic digests from four representative isolates of *B. cepacia*, *B. cenocepacia*, *B. multivorans*, and *B. gladioli*. Peptides are shown for *B. cepacia* (a), *B. cenocepacia* (b), *B. multivorans* (c), and for *B. gladioli* and *B. cepacia* peptide ADDAPEQGAPAEGGDERPR (d). The numbers given above each chromatogram are the rdotp/R-ratio values. Note for comparison that the total protein (including labeled peptides) injected into ESI-LC-MS was 50x that injected into the nano LC-MS (ChipCube QQQ) to adjust for the differences in sensitivity of the two instruments.
assay development isolates (Table S6) were tested on an Agilent 6495 ChipCube QQQ mass spectrometer. Analysis was performed using the Skyline package, and rdotp and R-ratio values were calculated. Rdotp data represent the normalized dot products of the light transition peak areas with the heavy (labeled peptide) transition peak areas. The R-ratio data represent the ratios of light transition peak areas with the heavy transition peak areas for quantitative calculation. Based on this analysis, rules were formulated as follows: (i) A peptide was called positive if the rdotp was $>0.95$ AND the R-ratio value was $\geq 0.5$. (ii) A peptide marker was called negative if the rdotp + R-ratio was $\leq 1.1$. (iii) Manual inspection was required for any peptide that was not automatically called as positive or negative. During manual inspection, removal of one compromised transition was allowed. (iv) Final species-level call was based on 2 or more species-specific peptides defining a positive species-level identification. Application of these rules to the sample development set yielded correct calls for 78/81 (96.3% accuracy with a 95% CI of 89.6 to 99.2%) individual peptides, and 439/447 correct negative calls (98.2% accuracy with a 95% CI of 96.5 to 99.2%). A set of 10/528 individual peptides (1.9%) required manual inspection, of which 3 were correctly called positive and 7 were correctly called negative. Final species-level identification following summing of individual peptide calls was correct for 33/33 isolates (100% accuracy with a 95% CI of 89.4 to 100%) in the test development set.

Following the initial test development on the Agilent 6495 ChipCube QQQ instrument, we adapted the assay to an Agilent 6460 ESI QQQ coupled with an Agilent 1290 Infinity II UHPLC system with a fast gradient of 3 min. This change in assay design improves speed and adapts the assay to a mass spectrometer more commonly available to the clinical lab. The 33 assay development isolates were run on the Agilent 6460 ESI QQQ to confirm optimal call rules for this instrument (data not shown). Based on these data, only a slight modification to the rules was required as follows: for the Agilent 6460 ESI QQQ, individual peptides are called positive for an rdotp of $>0.95$ AND an R-ratio value of $\leq 0.2$. The call rules for this instrument are otherwise the same as for the Agilent 6495 ChipCube QQQ.

**Analytical evaluation of MRM assay performance with blinded isolate set.** As a next step, we evaluated the fast gradient MRM assay on the Agilent 6460 ESI QQQ coupled with an Agilent 1290 Infinity II UHPLC system using a blinded set of 120 bacterial isolates (Table S1 and Table S6). Primary mass spectrometry data underwent manual boundary adjustment when necessary. Automatic call rules for individual peptides were then applied as described above for this instrument. Individual peptides marked for manual inspection were examined by three independent operators, which included examination of LC-MS chromatograms in Skyline software. After all individual peptides had been called either by automatic or manual methods, final species-level calls were made for all specimens.

Out of a total of 1,920 individual peptides, 272 were automatically called positive, 1,638 were automatically called negative, and only 10 required manual review. The calls made by all three independent operators were concordant for each of these 10 peptides. Final species-level assignments were applied to individual peptide calls, and results were unblinded at that point and compared with reference identifications. All 50/50 (100% accuracy with a 95% CI of 92.9 to 100%) negative controls were correctly called as negative (not \textit{B. cepacia}, \textit{B. cenocepacia}, \textit{B. multivorans}, or \textit{B. gladioli}). Accurate calls to the species level were made for 20/20 \textit{B. cenocepacia} (100% accuracy with a 95% CI of 83.2 to 100%); 15/15 \textit{B. cepacia} (100% accuracy with a 95% CI of 78.2 to 100%); 19/19 \textit{B. multivorans} (100% accuracy with a 95% CI of 82.4 to 100%); and 15/16 \textit{B. gladioli} (93.8% accuracy with a 95% CI of 69.8 to 99.8%). The single misidentification of \textit{B. gladioli} as a \textit{B. multivorans} was determined to be due to preanalytical error (mislabeling of a tube), and a repeat run from the original specimen performed in triplicate correctly identified this isolate as \textit{B. gladioli} (Fig. S4). Disregarding this preanalytical error, the accuracy and confidence interval for the mass spectrometry method itself would be 16/16 (100% accuracy.
with a 95% CI of 79.4 to 100%). Overall correct positive identifications, including the preanalytical error, were 69/70 (98.6% accuracy with a 95% CI of 92.3 to 99.96%), and 70/70 (100% accuracy with 95% CI of 94.9 to 100%) for the mass spectrometry method itself, excluding the preanalytical error.

Additional analysis of this data set demonstrated that boundary adjustment was important for peptide ADDAPEQGAPAEQGDERPR in B. cepacia to avoid false-positive detection of a similar peptide (ADDAPEQDAPAGEGDERPR) present in B. multivorans (Fig. S5). Both peptides have same m/z but slightly different retention times. Furthermore, it was found that the results of the 10 manual individual peptide calls did not have the potential to influence the final species-level calls, as they were fully determined by the other automatically called peptides for each identification; thus all final calls were determined automatically without operator interpretation other than initial boundary adjustments.

**DISCUSSION**

In this work, we have presented a general approach for the development of phylogeny-informed proteomic assays for the diagnostic classification of bacterial pathogens with reticulated phylogenies, using the B. cepacia complex (Bcc) as an example. The phylogeny of Bcc has significant network structure, which results in extensive phenotypic and molecular overlap that has complicated the construction of practical approaches for identification in the clinical laboratory. As a consequence, many clinical laboratories do not have capabilities to perform accurate species identification within the Bcc and rely instead on a small number of reference laboratories (23). MALDI-TOF mass spectrometry is now used routinely in hospital microbiology laboratories for rapid identification of many other bacterial pathogens, but commercial MALDI-TOF MS approaches, in which identification is based in part on conserved ribosomal and heat shock proteins, have difficulty differentiating species within the Bcc, similar to marker gene-based methods (13, 23, 24). Methods that take into account recombination are best suited to guide selection of optimal genetic markers for diagnostic assay development.

To gain an understanding of the Bcc phylogeny, we followed two approaches. In the first, we constructed coalescent trees using recombination-free markers. Comparison of these trees to those constructed without exclusion of recombination demonstrated significant differences in topology. We then pursued a second approach by constructing a network with explicit representation of recombination. To allow tractable computation, we approached this problem with a method that involved dividing the data set into blocks of ~200 assemblies/locus. These blocks were then assembled to include clades of interest that were selected based on pairwise distances with reference genomes, and each block was analyzed separately to produce a phylogenetic network. We then manually introduced boundaries separating most of the core network reticulations from the species-specific branches that were largely outside the core network. Placement of these boundaries was based on expert judgement and guided by an assumption that NCBI genome taxonomy contained some number of misclassifications, but that the majority of NCBI assignments were accurate. Genomes were then grouped with species by these boundaries and used as the basis for peptide selection. We note that an alternative approach could involve construction of a recombination-aware phylogeny with ClonalFrameML, as described above, followed directly by peptide selection outside areas affected by recombination.

The final MRM assay was based on 16 highly discriminatory peptides differentiating B. cepacia, B. cenocepacia, B. multivorans, and B. gladioli. The MRM assay was implemented with a set of positive and negative rules that were triggered automatically, with rules specifying when a spectrum required manual inspection. The assay was developed for the Agilent 6495 ChipCube QQQ instrument and then adapted to an Agilent 6460 ESI QQQ coupled with an Agilent 1290 Infinity II UHPLC system. The latter instrument setup allows a faster gradient of 3 min and represents more widely available technology in clinical laboratories. We estimate that with a standard setup, single...
samples can be run from start to finish in under 90 min, with per-sample run times decreasing significantly with batch size. The primary cost considerations involved in implementation of this assay are those pertaining to the mass spectrometry instrument itself and the operating personnel.

Testing with a blinded set of isolates on the Agilent 6460 ESI QQQ assay demonstrated excellent performance, with 50/50 correct automatic negative calls and 70/70 correct automatic species-level positive identifications after accounting for a single missed identification due to a preanalytic error, which was correctly identified on retesting. The accuracy of this approach compares favorably with many prior methods published in the literature (23). This approach is also more rapid that other approaches that require a combination of phenotypic testing and PCR. Importantly, our assay can be implemented on mass spectrometry instruments that are widely available in clinical laboratories and is amenable to automated interpretation. A potential challenge with the approach described here is that manual curation of the phylogenetic partitions relies on operator judgement. A fully automated solution to partitioning would be preferable and should be a focus of future development. In the case of the Bcc assay developed in this work, we would note that the partitions were validated by an expert clinical bacteriologist, and the peptides chosen based on these partitions yielded 100% analytical sensitivity and specificity for species-level identification in a blinded test set. We believe that these independent results confirm the accuracy of the manual curation approach used in this case.

In conclusion, the approach to analysis described here is applicable to many other pathogen groups for which development of diagnostic classification methods is complicated by interspecies recombination and network-structured phylogenies. We anticipate that with the rapidly increasing development of genomic databases for many clinically important pathogens, phylogeny-based proteomics approaches will have a progressively larger role to play in clinical microbiology, given the ease and speed of implementation and the precision of results obtainable.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.7 MB.

ACKNOWLEDGMENTS

This study was supported in part by the Intramural Research Programs of the National Institutes of Health Clinical Center, the National Institute of Allergy and Infectious Diseases, the National Heart, Lung and Blood Institute, and the National Institute of Diabetes and Digestive and Kidney Diseases. The BcRLR is supported by the Cystic Fibrosis Foundation. This study used the Office of Cyber Infrastructure and Computational Biology (OCICB) High Performance Computing (HPC) cluster at the National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, MD. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.


We declare no conflicts of interest.
Requests for materials should be addressed to John Dekker at john.dekker@nih.gov. Approval of requests may be subject to a Material Transfer Agreement.

REFERENCES