

METHODOLOGY ARTICLE

Open Access

A phylogenomic approach to bacterial subspecies classification: proof of concept in *Mycobacterium abscessus*

Joon Liang Tan^{3,4}, Tsung Fei Khang^{2*}, Yun Fong Ngeow³ and Siew Woh Choo^{1,4*}

Abstract

Background: *Mycobacterium abscessus* is a rapidly growing mycobacterium that is often associated with human infections. The taxonomy of this species has undergone several revisions and is still being debated. In this study, we sequenced the genomes of 12 *M. abscessus* strains and used phylogenomic analysis to perform subspecies classification.

Results: A data mining approach was used to rank and select informative genes based on the relative entropy metric for the construction of a phylogenetic tree. The resulting tree topology was similar to that generated using the concatenation of five classical housekeeping genes: *rpoB*, *hsp65*, *secA*, *recA* and *sodA*. Additional support for the reliability of the subspecies classification came from the analysis of *erm41* and *ITS* gene sequences, single nucleotide polymorphisms (SNPs)-based classification and strain clustering demonstrated by a variable number tandem repeat (VNTR) assay and a multilocus sequence analysis (MLSA). We subsequently found that the concatenation of a minimal set of three median-ranked genes: DNA polymerase III subunit alpha (*polC*), 4-hydroxy-2-ketovalerate aldolase (*Hoa*) and cell division protein FtsZ (*ftsZ*), is sufficient to recover the same tree topology. PCR assays designed specifically for these genes showed that all three genes could be amplified in the reference strain of *M. abscessus* ATCC 19977^T.

Conclusion: This study provides proof of concept that whole-genome sequence-based data mining approach can provide confirmatory evidence of the phylogenetic informativeness of existing markers, as well as lead to the discovery of a more economical and informative set of markers that produces similar subspecies classification in *M. abscessus*. The systematic procedure used in this study to choose the informative minimal set of gene markers can potentially be applied to species or subspecies classification of other bacteria.

Background

The rapidly growing mycobacteria (RGM), defined as mycobacteria that grow in culture media within seven days, are mostly environmental organisms, some of which have emerged as important opportunistic pathogens in humans, colonizers in the airway [1] or contaminants on surgical instruments [2]. More than a hundred species have so far been identified [3]. As different species and their subgroups are known to differ in their pathogenicity and susceptibility to antibiotics [4,5], their

rapid and accurate identification is important for making therapeutic decisions in patient management. Unfortunately, although many taxonomic advances have been made in recent years, resulting in the creation of new species, subspecies and subgroups, the taxonomic status of some members is still unresolved.

Mycobacterium abscessus is generally regarded as the most important RGM associated with human infections that range from localized cutaneous inflammation to rare but serious disseminated sepsis [6,7]. It is responsible for more than 80% of the chronic lung diseases

* Correspondence: tfkhang@um.edu.my; lchoo@um.edu.my

²Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603, Kuala Lumpur, Malaysia

¹Department of Oral Biology and Biomedical Sciences, Faculty of Dentistry, University of Malaya, 50603, Kuala Lumpur, Malaysia

Full list of author information is available at the end of the article

caused by RGM, some of which require surgical resection of the infected lung for complete resolution of symptoms [8]. This organism shows many similarities with another RGM, *M. chelonae*, and was once classified as *M. chelonae* subspecies *abscessus* [9]. With DNA-DNA hybridization studies, however, it became recognized as a distinct species [10,11]. Subsequently, the species was divided into three subspecies: *M. abscessus* sensu stricto, *M. massiliense* and *M. bolletii* [12-14] but the most recent classification describes just two subspecies which are *M. abscessus* subspecies *abscessus* (formerly *M. abscessus* sensu stricto) and *M. abscessus* subspecies *bolletii* that includes the previous two subspecies of *M. bolletii* and *M. massiliense* [15]. However, in many phylogenetic and phylogenomic studies, *M. abscessus* is clearly separated into three subgroups [16-18]. Clinical and population studies have also indicated notable differences among the three subspecies [16,17,19]. Hence, in this paper, we use the former three subspecies classification for ease of reference to earlier publications.

Traditionally, species identification for RGM is based on biological and biochemical tests such as pigment production, 3-day arylsulfatase reaction, nitrate reduction, iron uptake, and tolerance to 5% NaCl. Antibiotic susceptibility has also been used to assist in species or subspecies classification. For instance, polymyxin B inhibits *M. fortuitum* but not *M. abscessus* [20], while in the *M. abscessus* group, *M. massiliense* can be differentiated from the other two subspecies by a truncated erythromycin ribosome methyltransferase (*erm 41*) gene associated with susceptibility to macrolide antibiotics [19]. Although easy to use, the accuracy of these tests can be easily affected by the phenomenon of horizontal gene transfer (HGT) and differential gene expression. Therefore, phenotypic tests have largely been replaced by modern genotypic methods.

The PCR-Restriction Enzyme analysis (PRA) is one of the most commonly used genotypic methods. Direct sequencing of PCR amplified products based on the polymorphism of housekeeping genes has also become available in many diagnostic laboratories. In both methods, the *hsp65* gene, which is highly conserved within species, is frequently used. Most of the RGM currently known can be identified from RE patterns generated from a 439 bp variable portion of the gene [21,22]. In PCR-sequencing, the *hsp65* gene shows more variability among RGM species than the 16S rRNA gene, the universally conserved gene chosen for the phylogenetic analysis of prokaryotes. Thus, it is better than the latter for the identification of closely-related mycobacterial species [23]. Other housekeeping genes that have been used in RGM studies, albeit at lower frequency, include the internal transcribed spacer

(ITS) of the 16S-23S rRNA gene [24], the *rpoB* [25], *sodA* [26], *secA* [16], *recA* [27] and *gyrB* [28] genes. Commercially available reverse line probe hybridization assays (INNO-LIPA Mycobacteria, Innogenetics, Ghent, Belgium; GenoType Mycobacterium CM/AS, Hain Life-science GmbH, Germany) are increasingly being used as convenient molecular tools for rapid subspecies identification in diagnostic laboratories.

The conventional phylogenetic approach uses single or limited numbers of genes to infer phylogenetic relationships among the taxa of interest. Classification of taxa can then be done on the basis of the topology of the inferred phylogenetic tree. A potential drawback of this method is that the optimality of phylogenetic signal in the genes used, as quantified using some suitable metric, is largely unknown because there is no systematic procedure to find suitable candidates from a population of genes. The recent introduction of next-generation sequencing has made it possible to use whole-genome sequence information for subspecies classification via phylogenomic clustering. An example is the phylogenomic reconstruction of lactic acid bacteria by Zhang et al. [29]. The purpose of this study is to test the phylogenetic informativeness of five classical housekeeping genes: *rpoB*, *hsp65*, *secA*, *recA* and *sodA* that are commonly used in *M. abscessus* subspecies classification. Concurrently we wish to propose a phylogenomic approach to identify, to the subspecies level, 12 *M. abscessus* strains isolated from clinical samples, using a smaller set of genes that is at least as informative as the five classical genes for the purpose of inferring a high-confidence tree topology.

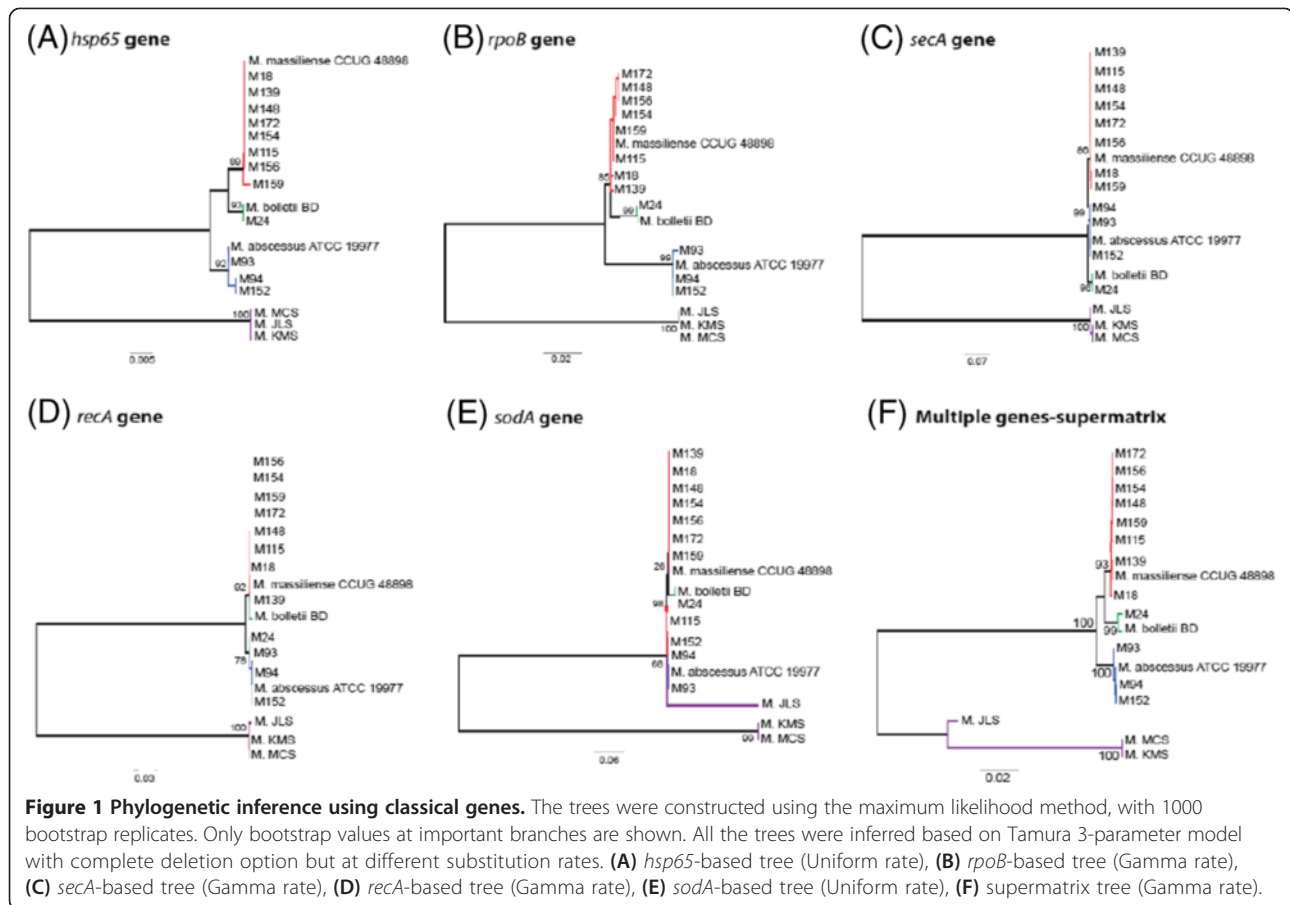
Results

Genome assembly, annotation and identification of orthologs

The draft genomes of our 12 *M. abscessus* strains, assembled using CLC Genomics Workbench and annotated using RAST, showed genome sizes ranging from 4,802,413 bp to 5,488,620 bp with an average of 5,054,711 bp. The number of putative Coding Sequences (CDS) ranged from 4,709 to 5560 (Additional file 1: Table S1). With the use of BLASTClust on the putative CDS from all genomes including those of the three reference strains and the three outgroup species, 210 orthologous genes were identified.

Single and multiple gene-based subspecies classification

On comparing the phylogenetic trees constructed using the five common marker genes for *M. abscessus* subspecies classification, the single gene-based trees inferred using *rpoB*, *hsp65* and *secA* genes showed clear and consistent classification of each isolate into subspecies (Figure 1A-C). The *recA* gene, however,



gave a slightly different classification (Figure 1D). For instance, the two reference strains, *M. bolletii* BD^T and *M. massiliense* CCUG 48898^T were grouped together and strain M24 was grouped with *M. abscessus* ATCC 19977^T, instead of *M. bolletii* BD^T as in the *rpoB*-, *hsp65*- and *secA*-derived trees. The *sodA*-based tree was also different in that strain M115 was identified as *M. abscessus* sensu stricto instead of *M. massiliense* as observed in the other trees (Figure 1E). These results indicate that *sodA* and *recA* marker genes, which are commonly used for mycobacterial classification, may not be suitable for *M. abscessus* subspecies identification.

The supermatrix tree obtained from the concatenation of the five genes has also revealed three clear subspecies grouping supported by strong bootstrap values of 93% for *M. massiliense*, 99% for *M. bolletii* and 100% for *M. abscessus* sensu stricto (Figure 1F). The classification for our 12 strains was similar to those obtained with the *rpoB*, *hsp65* and *secA* genes.

Phylogenomic approach to subspecies classification

The availability of whole-genome sequences enabled us to systematically evaluate the phylogenetic informativeness

of a large pool of candidate genes and then select optimal ones for phylogenetic tree construction. For this evaluation, we used the same procedure to predict the Open Reading Frames (ORFs) in the 12 genomes sequenced in this study, as well as the genomes of the reference and outgroup species. On aligning the sequences of the 210 orthologous genes identified with BLASTClust, we found the MSA produced using MAFFT [30] to be more robust than that produced using MUSCLE [31]. Subsequently, we chose the MAFFT-aligned MSA for downstream analysis. The phylogenetic informativeness of each gene was assessed and ranked based on calculated entropy values as described in the Methods section. By using this systematic approach, we selected 50 median-ranked genes which would be the most phylogenetically optimal for the purpose of subspecies phylogeny inference. The functional analysis of these 50 median-ranked genes, as well as the 50 top-ranked and 50 bottom ranked genes showed that the majority of the bottom-ranked genes are involved in translational processes, whereas those of the top-ranked genes are mostly involved in enzyme metabolism (Figure 2; Additional file 1: Table S2–S4). These results indicate that the ranking

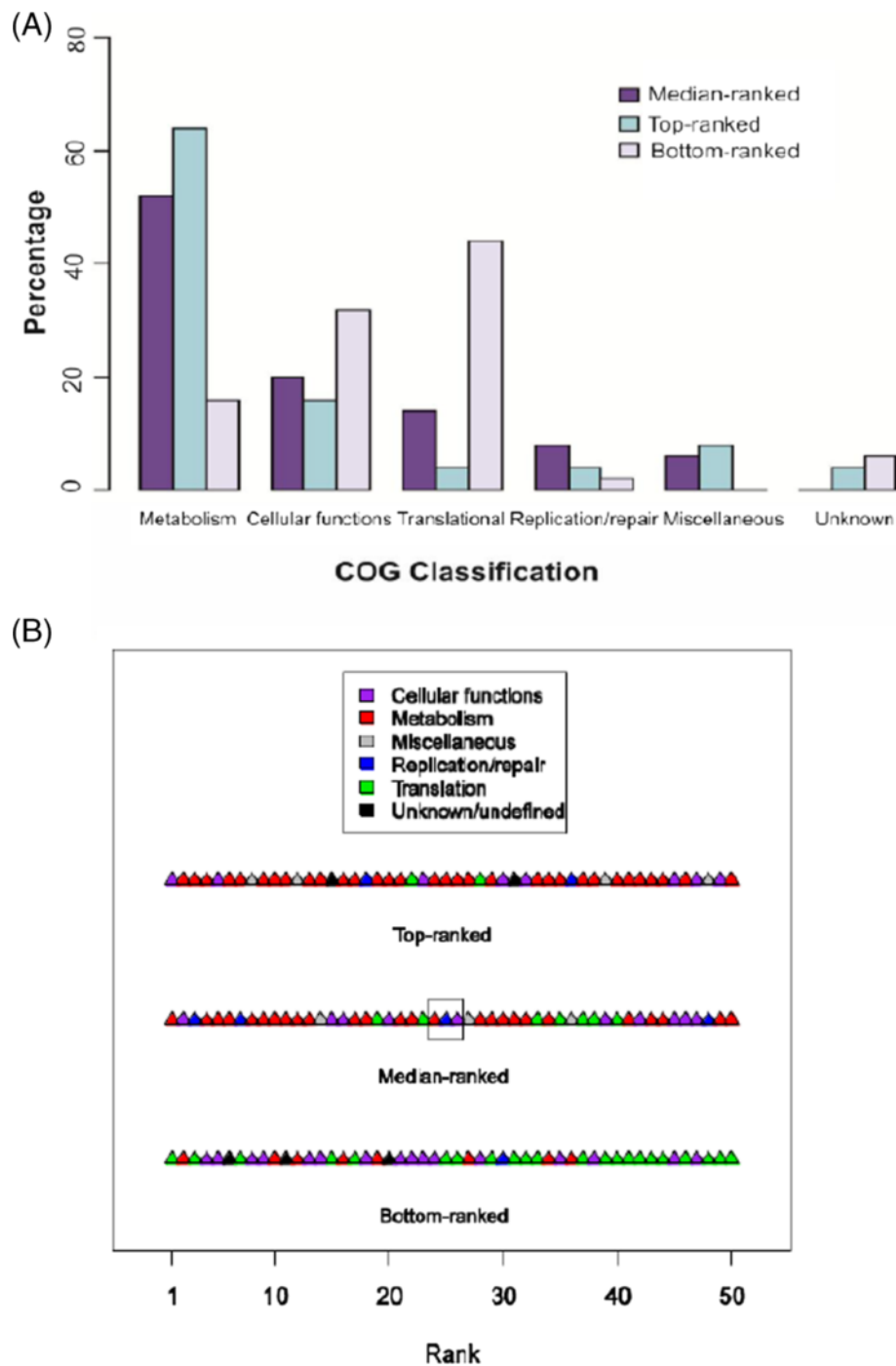
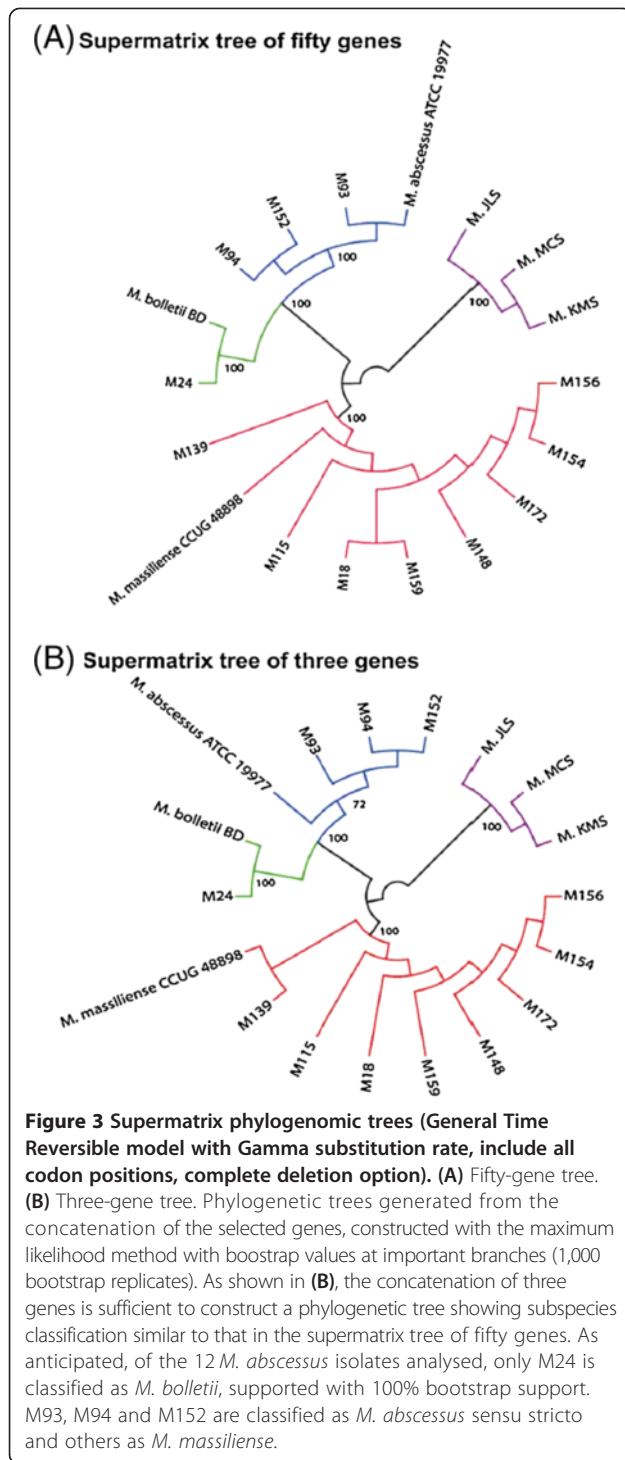


Figure 2 Functional analysis of orthologs. (A) COG classification composition of genes in the top, median and bottom-ranked genes. **(B)** COG classification of genes in the top, median and bottom-ranked genes. The highest relative entropy is found in the left-most member of the top-ranked genes; the lowest entropy is found in the right-most member of the bottom-ranked genes. The set of minimal genes capable of reconstructing the clustering topology as the one obtained using 50 median-ranked genes is boxed.

obtained using the relative entropy metric appears to be biologically reasonable.

In the phylogenomic tree constructed based on the selected 50 median-ranked genes (Figure 3A), the classification

of isolates inferred agreed with the one inferred using the five concatenated classical housekeeping genes. Strains M93 to M94 were clustered with *M. abscessus sensu stricto*; M24 with *M. bolletii* and the others with *M. massiliense*.



To find a minimal set of genes capable of producing the same subspecies classification as that obtained using the 50 median-ranked genes, we iteratively removed genes from the left and right ends of the list of median-ranked genes, and used the remaining genes to construct the phylogenetic tree. The process was terminated

when the classification implied by the tree topology changed. Thus, we found that the minimum number of median-ranked genes needed for recovering similar subspecies classification (Figure 3B) attained using the 50 median-ranked genes was three. These three genes were DNA polymerase III subunit alpha (*polC*; Accession: YP_001703430.1), 4-hydroxy-2-ketovaleate aldolase (*Hoa*; Accession: YP_001701378.1) and cell division protein FtsZ (*ftsZ*; Accession: HQ662067.1). We confirmed the amplifiability of informative fragments from these three genes in *M. abscessus* ATCC 19977^T (Additional file 1: Figure S1; Additional file 1: Table S5).

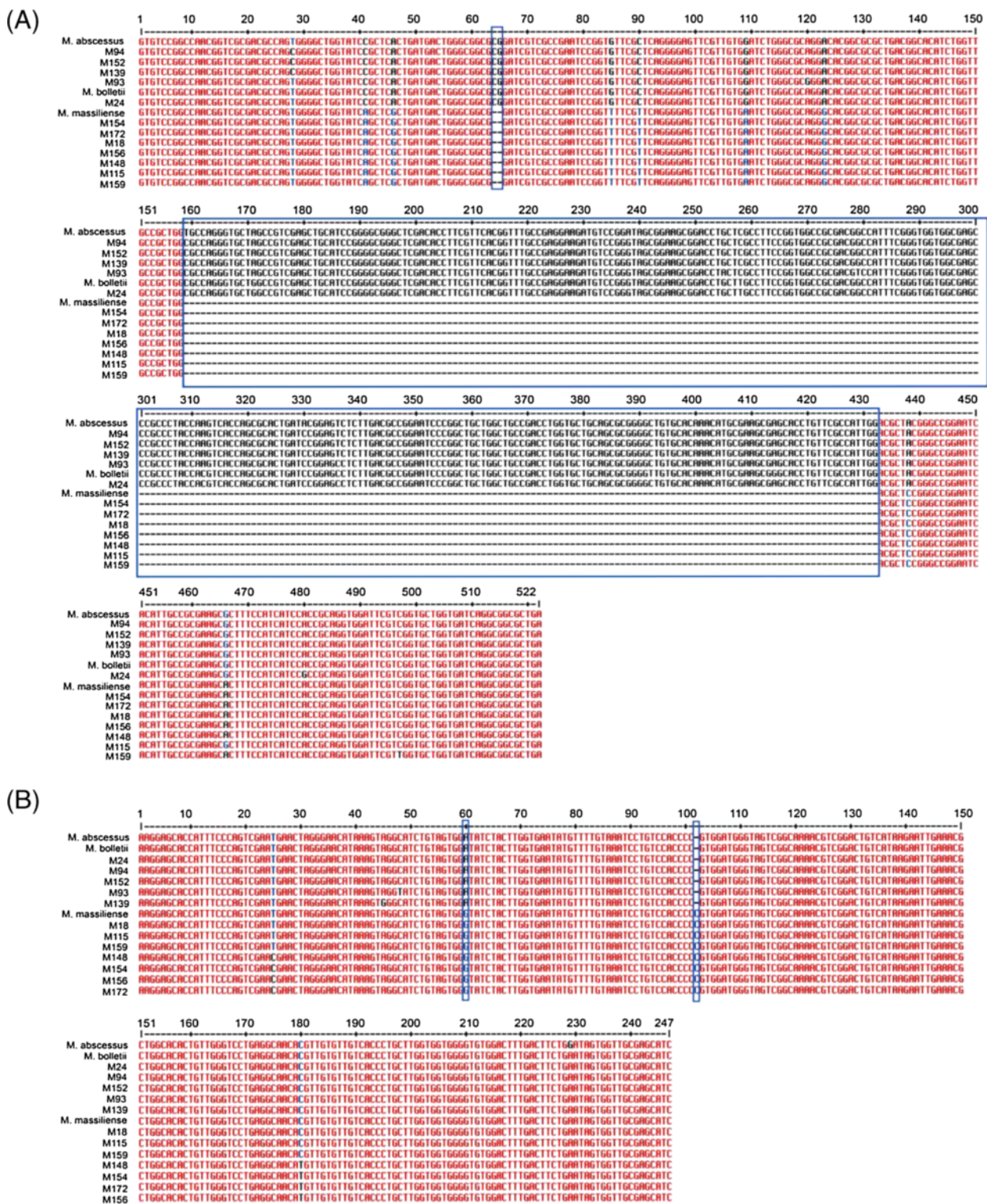
Evidence to support the subspecies classification by phylogenomics analysis

Examination of *erm41* and 16S-23S rRNA ITS genes

Previous studies have shown that the *erm41* and the 16S-23S rRNA ITS genes could be used to differentiate the subspecies of *M. abscessus* [12,19]. We examined the sequences of these genes in all 12 strains and found seven strains (M156, M148, M154, M172, M152, M115 and M18) having deletions at the 64th and 65th positions and a large deletion from position 159th to 432nd in the *erm41* gene (Figure 4A), as well as an A to G substitution at position 60th and insertion of C at position 102nd in the ITS gene (Figure 4B). Since these features are typically found in *M. massiliense*, their presence in the seven strains further supported the classification of these strains as *M. massiliense*, as observed in the supermatrix analyses. However, M139 which was classified as *M. massiliense* in all the single gene-based and multiple gene-based approaches did not have these expected features.

Classification of other *M. abscessus* strains from public database

At the end of our study, many *M. abscessus* genome sequences have been deposited in NCBI Genbank database. To further test the reliability of the three selected gene markers, we used them to classify 43 *M. abscessus* genomes (12 from our laboratory; 28 deposited by other researchers in the NCBI Genbank database; three reference) and three outgroup genomes (Additional file 1: Table S6). The phylogenetic tree constructed using the minimal set of three genes clearly showed three major groups corresponding to the three *M. abscessus* subspecies (Figure 5). For example, the strains from 6G0125S to 4S0726RB were grouped into the *M. abscessus* sensu stricto cluster, whereas the strains from 47J26 to 5S0422 were likely to be *M. massiliense*. None of the additional 28 strains were identified as *M. bolletii*. This classification was in accordance with and supported by the analysis of *hsp65* and *rpoB* gene sequences (Additional file 1: Figure S2). The addition of 28 strains into this study



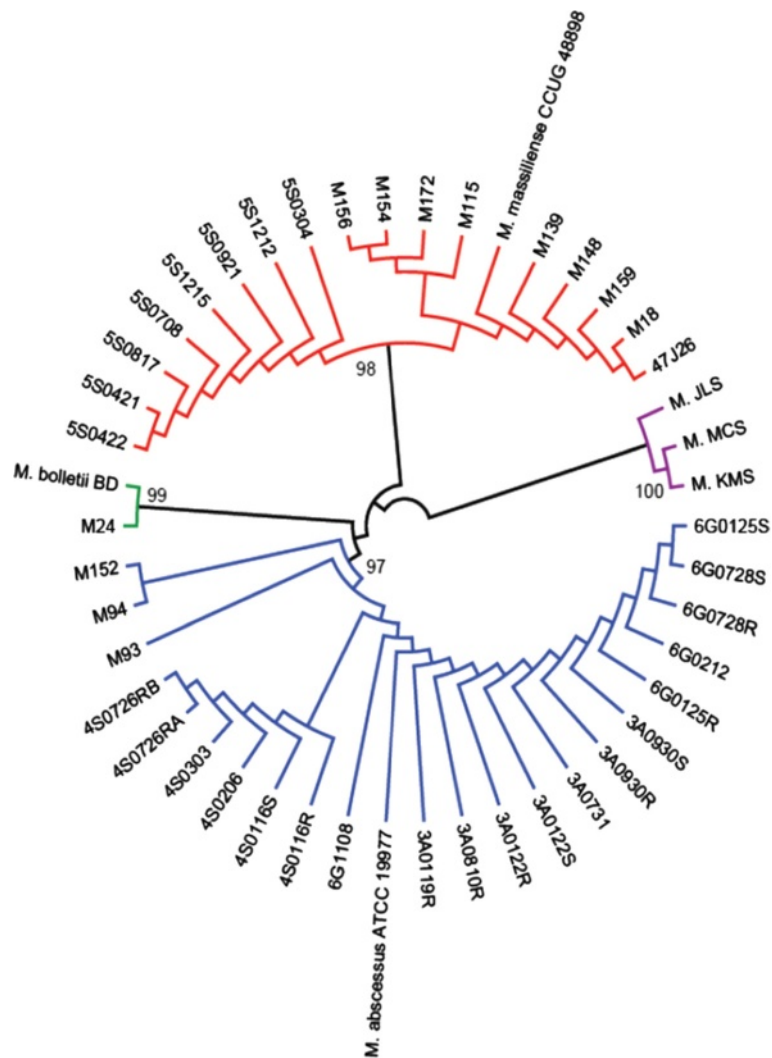


Figure 5 Phylogenetic tree based on three marker genes, for 43 *M. abscessus* strains and three outgroups (Tamura 3-parameter model with Gamma substitution rate, include all codon positions, complete deletion option). Of the 28 additional *M. abscessus* strains, this phylogenomics analysis revealed 19 strains as *M. abscessus* sensu stricto and 9 strains as *M. massiliense*, supported by 97% to 98% bootstrap support.

did not change the classification of the 12 strains from our initial analysis. These results again support our belief that the three selected genes are reliable markers for the classification of *M. abscessus* subspecies.

Other supporting evidence

Recently, Wong *et al.* [32] showed that a variable-number tandem-repeat (VNTR) typing assay for the *M. abscessus* species seemed to give good differentiation of *M. abscessus* strains. In their study, both the VNTR assay and a multilocus sequence analysis assay (MLSA) based on seven housekeeping genes (*argH*, *cya*, *glpK*, *gnd*, *murC*, *pta*, and *purH*) yielded subspecies

clustering that is concordant (with the exception of M139) with the one inferred using the present phylogenomics approach.

In addition, the availability of whole genome sequences of the 12 *M. abscessus* strains allowed us to proceed with the reconstruction of a phylogenetic tree using single nucleotide polymorphisms (SNPs) in the genomic regions that are conserved across all strains (Additional file 1: Figure S3). The resulting SNP-based tree showed that the classification of the 12 *M. abscessus* strains is similar to the classification obtained from the phylogenomics tree generated using our systemic data mining approach.

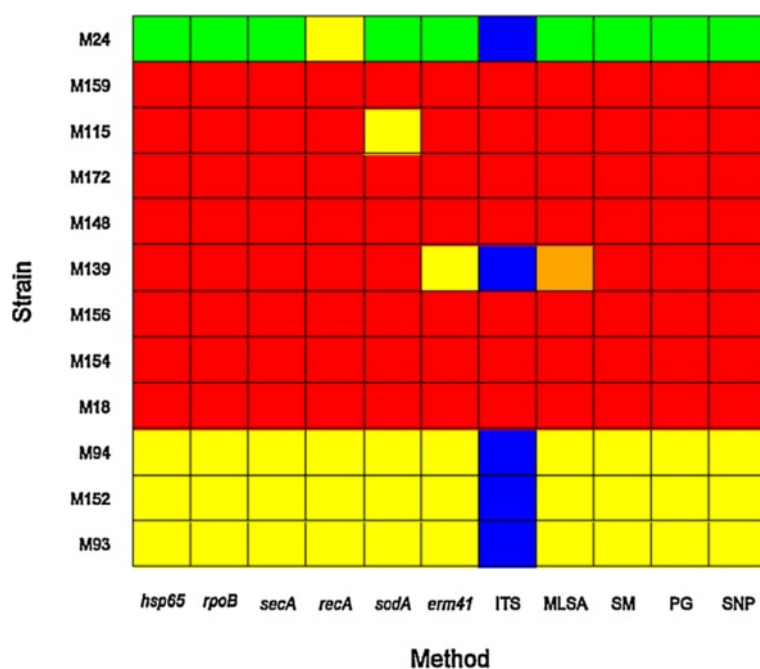


Figure 6 Classification of strains according to different methods. Color legend: Yellow for *M. abscessus sensu stricto*; red for *M. massiliense*; green for *M. bolletii*; blue for *M. abscessus sensu stricto/M. bolletii*; orange for *M. abscessus sensu stricto/M. massiliense*. Abbreviation: SM: supermatrix; PG: phylogenomics; SNP: single-nucleotide polymorphism.

The classification results of the 12 sequenced genomes using different approaches are summarized in Figure 6.

Discussion

Using a proposed phylogenomic approach, we have successfully classified into subspecies, 12 sequenced genomes of *M. abscessus* strains isolated from clinical samples. The primary value of this approach is the use of a set of objectively selected genes for constructing a high confidence tree topology that can serve as a benchmark. Subsequently, the classification implicit in the topology of the inferred tree can be used to guide the selection of a minimal gene set for practical, routine use. Existing methods (multiple-genes, MLSA) are not able to produce such benchmarks, and their resulting classification accuracy may be largely a matter of luck rather than science. Corroboration from the analysis of whole-genome SNP data further convinces us of the usefulness of the phylogenomic approach to systematically discover a minimal gene set for subspecies classification, at least in *M. abscessus*.

Our phylogenomic approach uncovered a set of genes detected using median-ranking that supports the subspecies classification inferred using the concatenation of the five classical genes. The set of median-ranked genes has good coverage of all important functional classes compared to the top and bottom-ranked genes, so using

their concatenation for phylogenetic inference potentially avoids systematic biases that can be caused by having a concentration of genes in the ranked list for some particular functional class. Further refinement resulted in a new, three-member gene set that preserves the similar classification. The three selected genes are biologically interesting and come from three different functional classes. The KEGG database [33] shows that DNA polymerase III subunit alpha (*polC*) functions in DNA replication; 4-hydroxy-2-ketovaleate aldolase is involved in phenylalanine metabolism, degradation of benzoate, dioxin and xylene; and cell division protein gene *ftsZ* produces the FtsZ GTP-binding cell division protein which is important for the cytoskeleton formation during cell wall synthesis. A brief survey of available literature shows that *polC* has been used as a phylogenetic marker in *Bacillus subtilis* subspecies classification [34] as well as lactic acid bacteria phylogeny [29]. In *Wolbachia* phylogenetic studies, *ftsZ* has been used in conjunction with other genes [35,36].

In this study, we also assessed a set of genes commonly used by other workers for the subspecies classification of *M. abscessus* (*rpoB*, *hsp65*, *recA*, *secA*, *sodA*, *erm41* and *ITS*) and found inconsistent classification for certain strains with *recA*, *sodA*, *erm41* and *ITS* genes. It is well-known that the single gene-based approach is sensitive to the evolutionary history of the

gene and not necessarily of the species, and may be limited by the lack of sufficient variation in a single gene sequence [37-39]. A multiple gene-based approach, if the genes are carefully chosen, may amplify the individual phylogenetic signals present in the genes and lead to an accurate inference of phylogenetic tree for subspecies classification [40,41]. This is seen in the superior robustness of our supermatrix trees whether they are based on the concatenation of five classical genes, core genome SNPs or our three gene markers.

Of the 12 strains we isolated, M139 had the most ambiguous taxonomic identity. It was classified as *M. massiliense* in the current study and also in two recent publications [17,18] using our *M. abscessus* genome sequences for comparative analyses. However, it had features of *M. abscessus sensu stricto* in the *erm41* and *ITS* genes and was shown to cluster with *M. abscessus* in the MLS and VNTR assays reported by Wong *et al.* [32]. Additionally, it was identified as *M. abscessus* subsp. *abscessus* in the MALDI Biotyper system (Bruker, Germany) that uses the MALDI-TOF (matrix-assisted laser desorption ionization-time of flight) mass spectrometry to identify organisms by their molecular (protein) fingerprints. The *erm41* gene in M139 is intact but the minimum inhibitory concentration (MIC) of clarithromycin obtained with the Epsilometer test (bio Merieux, France) was 0.094 mg/l even after prolonged incubation up to 14 days. This high level of *in vitro* susceptibility is more often seen among *M. massiliense* than *M. abscessus* or *M. bolletii* strains. In M139, however, this low MIC can be attributed to the C28 polymorphism in the *erm41* gene which has been reported to prevent the expression of the inducible clarithromycin resistance conferred by an intact *erm41* gene [42].

The apparent incongruities in the subspecies classification of M139 could be because previous studies used insufficient numbers of strains to describe "typical" features associated with the different *M. abscessus* subspecies. More extensive sampling of *M. abscessus* populations from different geographical locations and clinical settings would help to resolve current taxonomic uncertainties.

Conclusion

Whole-genome sequencing of medically important bacterial strains opens up new possibility for accurate subspecies classification via a phylogenomic approach. We have shown how a large pool of orthologs can be identified bioinformatically from the genomes of 12 *M. abscessus* isolates, assembled from fragments generated using Illumina shotgun sequencing. We then introduced a method to rank the phylogenetic informativeness of the identified orthologs using the relative entropy metric. Based on the set of 50 median-ranked

genes, a benchmark phylogeny was obtained, and we used its topology to infer subspecies classification. We made the method practical by discovering a three-member minimal gene set that could return the same subspecies classification obtained using 50 median-ranked genes. The workflow here provides an objective means for the development of molecular-based classification method that was not possible with previous methods such as single gene, arbitrary multiple gene approaches, VNTR as well as MLSA. Our three gene set is at least as informative as classical markers commonly used by other researchers. Its applicability can be further evaluated by testing it on a larger population of strains, from different clinical settings and geographical locations.

Methods

Bacterial strains

The 12 *M. abscessus* strains used in this study were collected between July 2009 and June 2011 from the Clinical Microbiology Laboratory of University of Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia. With the exception of one strain from a lymph node, all other strains were from sputum or bronchoalveolar fluid of patients presenting with respiratory infections. In the UMMC laboratory, clinical specimens from patients suspected to have tuberculosis are routinely cultured in the BACTEC MGIT₉₆₀ liquid culture system (Becton-Dickinson). Positive cultures are examined microscopically and subcultured on Lowenstein-Jensen (LJ) slants. Acid-fast colonies on LJ slants are then identified using a variety of tests. The strains used in this study were identified as *M. abscessus* by their rapid growth (within seven days of incubation) in the MGIT₉₆₀ system, non-pigmented acid-fast colonies on LJ slants, typical restriction pattern in a PCR-restriction fragment length polymorphism analysis [21] and a positive identification in a reverse line probe hybridization assay (GenoType Mycobacterium CM/AS; Hain Life-science GmbH, Germany).

All isolates were kept in Middlebrook 7H9 broth with 15% glycerol, at -80°C, until required for further testing.

For whole-genome sequencing, archived strains were retrieved, rendered non-nominal and subcultured on LJ slants. This study involved only genomic analysis of routine isolates and, with the exception of specimen type and geographical origin, none of the data used could expose patient identity. As such, we considered it unnecessary to apply for ethical approval by the University's Medical Ethics Committee Standard Operating Procedures (<http://www.ummc.edu.my/view/content.php?ID=56>).

The genome information of the reference strains was extracted from the NCBI Genbank [43] database. The

reference strains used for comparison are given in Additional file 1: Table S6.

DNA extraction & isolate validation

Pure cultures on LJ slants were harvested by flooding each slant with 3 ml of phosphate buffered saline (PBS, Ph 7.4) followed by gentle scraping of the agar surface with a glass Pasteur pipette. The suspension was then collected into a 15 mL centrifuge tube and vortex mixed for 5 minutes. For DNA extraction, 200 μ L of suspension was processed with the ZR Fungal/Bacterial DNA MiniPrep™ (Zymo Research, USA) according to the protocol provided by the manufacturer. DNA concentration and purity were measured on the NanoDrop2000 spectrophotometer (Thermo Scientific). To validate the identity of the 12 *M. abscessus* isolates, we used PCR to amplify their *hsp65* genes for Sanger sequencing, and then used BLAST to query the sequences to known *Mycobacterium* species/subspecies gene sequences stored in a web-accessible database of *hsp65* locus sequences [44]. All 12 strains achieved almost perfect or perfect (98% to 100%) similarity to *M. abscessus*.

Whole-genome sequencing, annotation and identification of orthologs

The genomes of the 12 isolates were shotgun-sequenced using Illumina Genome Analyzer 2X technology. The raw sequencing reads were trimmed at a threshold of 0.01 and the sequences obtained were assembled *de novo* using CLC Genomics Workbench version 4.9 (CLC bio, Denmark). All target genes for analysis were extracted from the assembled genomes of the 12 *M. abscessus* strains. The gene sequences for the reference *M. abscessus* subspecies were obtained from the NCBI database. We used BLAST [45] to locate the boundaries of the genes in the 12 assembled draft genomes.

We annotated the genomes of strains included in this analysis using Rapid Annotation running on Subsystem Technology (RAST) automated pipeline [46]. The orthologs among the putative open reading frames (ORFs) were identified using BLASTClust in a standalone BLAST program [47].

Phylogenetic classification using classical markers

The pan-bacterial 16S rRNA gene, which is generally considered the best target for bacterial identification and phylogenetic analysis, was not used in this study as it has been reported to lack discriminatory power for *M. abscessus* subspecies [48]. Instead, five other genes that have been used by other workers were chosen for phylogenetic analysis, namely, the *rpoB* (711 bp) gene encoding for the β -subunit of RNA polymerase enzyme responsible for RNA synthesis [49]; the *hsp65* (603 bp) gene for the 65 kDa heat shock protein that is involved in intracellular protein folding, assembly and transport [50]; the *sodA* (539 bp) gene

encoding superoxide dismutase, a metalloenzyme that contributes to defence against oxidative stress in mycobacteria [51]; the *recA* (1041 bp) recombinase gene related to DNA repair [52] and the *secA* (700 bp) gene that is involved with pre-protein transport in the bacterial cell [53].

Each of the five housekeeping gene sequences was aligned using ClustalX from the European Bioinformatics Institute (EBI) online server to study similarities and variable sites among the subspecies [54]. We then used MEGA version 5.10 [55] to select the best DNA substitution model for each multiple sequence alignment (MSA), followed by maximum likelihood estimation of trees. In addition, we concatenated all five genes, and similarly used the MSA of the latter to infer maximum likelihood trees (1,000 bootstrap replications).

Two additional genes, *erm41* and 16S-23S rRNA *ITS* with known features for subspecies identification were used to further support the classification results. *M. massiliense* can be recognized by the presence of a truncated *erm41* gene [19] as well as an A to G substitution at the 60th position and a C insertion at the 102nd position of the 16S-23S rRNA *ITS* [12].

Phylogenomic classification using whole-genome data

Not all genes have the same phylogenetic signal [56]. The main advantage of the phylogenomic approach is that it allows the generation of a large population of orthologs in the taxa of interest so that an objective assessment of their phylogenetic informativeness can be made. These genes can be ranked accordingly based on some metric that is entropy-dependent for the purpose of quantifying phylogenetic informativeness. Given a list of genes ranked from high to low entropy, we propose that a subset of genes in the neighbourhood of the median of this ranked list would be the most phylogenetically optimal for the purpose of subspecies phylogeny inference. If this method works, a simple functional categorization of genes in the top, median and bottom-ranked genes should roughly show the following pattern: dominance of genes involved in the highly conserved translational process in the bottom-ranked genes, and dominance of genes involved in metabolism in the top-ranked genes. The subsequent phylogenetic tree constructed using the median-ranked genes would then constitute a high confidence phylogeny.

To prepare the genes for ranking, we performed multiple sequence alignment (MSA) for the orthologs identified. Next, we removed all MSA with alignment gaps greater than 5% of the size of the alignment block. To reduce the chances of using genes that are not robust to choice of the MSA algorithm, we compared the results of MUSCLE and MAFFT. We chose these two algorithms because of positive evaluations with respect to their accuracy and speed performance [57]. To exclude co-optimality of MSA as a potential source of alignment error, we further performed

the Heads or Tails (HoT) analysis [58] on the identified orthologs. In HoT analysis, an MSA using reversed sequences was first built; this MSA was then reversed and then compared against the MSA built using the original, unreversed sequences. Only genes that show identical MSA after an HoT analysis were subsequently used for downstream analysis.

To quantify the phylogenetic signal of the i th gene (g_i) gene, we used the relative entropy metric for a gene's MSA:

$$H(g_i) = \sum_{x=1}^s \sum_{y=1}^L p_{xy}(nt) \log \frac{p_{xy}(nt)}{p_x(nt)p_y(nt)},$$

where s is the number of sequences, L is the length of the MSA, $p_{xy}(nt)$ is the relative frequency of state nt (gaps are treated as a fifth character) in the entire MSA at row x and column y , with $nt \in \{A, G, C, T, -\}$, $p_x(nt)$ is the row marginal frequency for state nt , and $p_y(nt)$ is the column marginal frequency for state nt . This metric measures the divergence between the nucleotide frequencies in the entire MSA from the nucleotide frequencies in the rows and columns of the MSA. An MSA with noiseless columns would yield a zero value since the nucleotide frequencies in the columns would be degenerate, and those in the rows would be the same as the nucleotide frequencies in the MSA. For MSA with noisy columns, H increases. All computation was done using an in-house R script [59] (available upon request).

Fifty genes - 25 to the left and right (in alternating order) of the list of median-ranked genes, were concatenated and used to reconstruct the phylogenetic tree of the 12 isolate strains. Taking the clustering of strains based on the 50 median-ranked genes as the gold standard, we subsequently deduced the set of minimum median-ranked genes needed to recover the same clustering by iteratively reducing the number of median-ranked genes used and checking the resulting change in clustering topology. Candidate markers in this minimum set were then tested for their amplifiability using a standard PCR protocol to establish their practical usefulness as new, objectively inferred phylogenetic markers.

Additional file

Additional file 1: Supplementary file.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

Conceived and designed the experiments: all authors. Performed the wet laboratory experiments and contributed reagents/materials: YFN. Analyzed the data: JLT, SWC and TFK. Wrote the paper: all authors. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Hien Fuh Ng and Yan Ling Wong for helping us to design primers and optimizing the PCR for the 3 new gene markers in this project.

This work was supported by the University of Malaya [Postgraduate Research Grant PG077-2012B] and Ministry of Higher Education, Malaysia [grant numbers: UM.C/HIR/MOHE/08 and LR004/2011A].

Author details

¹Department of Oral Biology and Biomedical Sciences, Faculty of Dentistry, University of Malaya, 50603, Kuala Lumpur, Malaysia. ²Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603, Kuala Lumpur, Malaysia. ³Department of Medical Microbiology, Faculty of Medicine, University of Malaya, 50603, Kuala Lumpur, Malaysia. ⁴Genome Informatics Research Laboratory, High Impact Research (HIR) Building, University of Malaya, 50603, Kuala Lumpur, Malaysia.

Received: 17 July 2013 Accepted: 10 December 2013

Published: 13 December 2013

References

1. Burns DN, Wallace RJ Jr, Schultz ME, Zhang YS, Zubairi SQ, Pang YJ, Gibert CL, Brown BA, Noel ES, Gordin FM: Nosocomial outbreak of respiratory tract colonization with *Mycobacterium fortuitum*: demonstration of the usefulness of pulsed-field gel electrophoresis in an epidemiologic investigation. *Am Rev Respir Dis* 1991, **144**:1153-9.
2. Wallace RJ Jr, Brown BA, Griffith DE: Nosocomial outbreaks/pseudo-outbreaks caused by nontuberculous mycobacteria. *Annu Rev Microbiol* 1998, **52**:453-90.
3. Schinsky MF, Morey RE, Steigerwalt AG, Douglas MP, Wilson RW, Floyd MM, Butler WR, Daneshvar MI, Brown-Elliott BA, Wallace RJ Jr, McNeil MM, Brenner DJ, Brown JM: Taxonomic variation in the *Mycobacterium fortuitum* third biovariant complex: description of *Mycobacterium boenickei* sp. nov., *Mycobacterium houstonense* sp. nov., *Mycobacterium neworleansense* sp. nov. and *Mycobacterium brisbanense* sp. nov. and recognition of *Mycobacterium porcinum* from human clinical isolates. *Int J Syst Evol Microbiol* 2004, **54**:1653-1667.
4. Brown BA, Springer B, Steingrube VA, Wilson RW, Pfyffer GE, Garcia MJ, Menendez MC, Rodriguez-Salgado B, Jost KC Jr, Chiu SH, Onyi GO, Böttger EC, Wallace RJ Jr: *Mycobacterium wolinskyi* sp. nov. and *Mycobacterium goodii* sp. nov., two new rapidly growing species related to *Mycobacterium smegmatis* and associated with human wound infections: a cooperative study from the International Working Group on Mycobacterial Taxonomy. *Int J Syst Bacteriol* 1999, **49**:1493-1511.
5. Yakrus MA, Hernandez SM, Floyd MM, Sikes D, Butler WR, Metchock B: Comparison of Methods for Identification of *Mycobacterium abscessus* and *M. chelonae* Isolates. *J Clin Microbiol* 2001, **39**:4103-4110.
6. Ingram CW, Tanner DC, Durack DT, Kernodle GW Jr, Corey GR: Disseminated infection with rapidly growing mycobacteria. *Clin Infect Dis* 1993, **16**:463-71.
7. Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, Holland SM, Horsburgh R, Huit G, Lademaro MF, Iseman M, Olivier K, Ruoss S, Von Reyn CF, Wallace RJ Jr, Winthrop K: ATS Mycobacterial Diseases Subcommittee: American Thoracic Society; Infectious Disease Society of America: An Official ATS/IDSA Statement: Diagnosis, Treatment, and Prevention of Nontuberculous Mycobacterial Diseases. *Am J Respir Crit Care Med* 2007, **175**:367-416.
8. Griffith DE, Girard WM, Wallace RJ Jr: Clinical feature of pulmonary disease caused by rapidly growing mycobacteria. An analysis of 154 patients. *Am Rev Respir Dis* 1993, **147**:1271-1278.
9. Kubica GP, Baess I, Gordon RE, Jenkins PA, Kwapinski JB, McDermont C, Pattyn SR, Saito H, Silcox V, Stanford JL, Takeya K, Tsukamura M: A co-operative numerical analysis of rapidly growing mycobacteria. *J Gen Microbiol* 1972, **73**:55-70.
10. Le'vy-Fre'bault V, Grimont F, Grimont PAD, David HL: Deoxyribonucleic acid relatedness study of the *Mycobacterium fortuitum*-*Mycobacterium chelonae* complex. *Int J Syst Bacteriol* 1986, **36**:458-460.
11. Kusunoki S, Ezaki T: Proposal of *Mycobacterium peregrinum* sp. nov., nom. rev. and elevation of *Mycobacterium chelonae* subsp. *abscessus* (Kubica et al.) to species status: *Mycobacterium abscessus* comb. nov. *Int J Syst Bacteriol* 1992, **42**:240-245.
12. Adékambi T, Reynaud-Gaubert M, Greub G, Gevaudan MJ, La Scola B, Raoult D, Drancourt M: Amoebal Coculture of "*Mycobacterium massiliense*" sp. nov. from the Sputum of a Patient with Hemoptoic Pneumonia. *J Clin Microbiol* 2004, **42**:5493-5501.

13. Adékambi T, Berger P, Raoult D, Drancourt M: *rpoB* gene sequence-based characterization of emerging non-tuberculous mycobacteria with descriptions of *Mycobacterium bolletii* sp. nov., *Mycobacterium phocaicum* sp. nov. and *Mycobacterium aubagnense* sp. nov. *Int J Syst Evol Microbiol* 2006, **56**:133–143.
14. Macheras E, Roux A-L, Ripoll F, Sivadon-Tardy V, Gutierrez C, Gaillard J-L, Heym B: Inaccuracy of Single-Target Sequencing for Discriminating Species of the *Mycobacterium abscessus* Group. *J Clin Microbiol* 2009, **47**:2596–2600.
15. Leao SC, Tortoli E, Euzéby JP, Garcia MJ: Proposal that *Mycobacterium massiliense* and *Mycobacterium bolletii* be united and reclassified as *Mycobacterium abscessus* subsp. *bolletii* comb. nov., designation of *Mycobacterium abscessus* subsp. *abscessus* subsp. nov. and emended description of *Mycobacterium abscessus*. *Int J Syst Evol Microbiol* 2011, **61**:2311–2313.
16. Zelazny AM, Root JM, Shea YR, Colombo RE, Shamputa IC, Stock F, Conlan S, McNulty S, Brown-Elliott BA, Wallace RJJ, Olivier KN, Holland SM, Sampaio EP: Cohort Study of Molecular Identification and Typing of *Mycobacterium abscessus*, *Mycobacterium massiliense*, and *Mycobacterium bolletii*. *J Clin Microbiol* 2009, **47**:1985–1995.
17. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA: Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet* 2013, **381**:1551–60.
18. Davidson RM, Hasan NA, de Moura VC, Duarte RS, Jackson M, Strong M: Phylogenomics of Brazilian epidemic isolates of *Mycobacterium abscessus* subsp. *bolletii* reveals relationships of global outbreak strains. *Infect Genet Evol* 2013, **20**:292–297.
19. Kim H-Y, Kim BJ, Kook Y, Yun Y-J, Shin JH, Kim B-J, Kook Y-H: *Mycobacterium massiliense* is differentiated from *Mycobacterium abscessus* and *Mycobacterium bolletii* by erythromycin ribosom methyltransferase gene (*erm*) and clarithromycin susceptibility patterns. *Microbiol Immunol* 2010, **54**:347–353.
20. Wallace RJJ, Swenson JM, Silcox VA, Good RC: Disk diffusion testing with polymyxin and amikacin for differentiation of *Mycobacterium fortuitum* and *Mycobacterium chelonae*. *Clin Microbiol* 1982, **16**:1003–1006.
21. Telenti A, Marchesi F, Balz M, Bally F, Böttger EC, Bodmer T: Rapid identification of mycobacteria to the species level by polymerase chain reaction and restriction enzyme analysis. *J Clin Microbiol* 1993, **31**:175–178.
22. Steingrube VA, Gibson JL, Brown BA, Zhang Y, Wilson RW, Rajagopalan M, Wallace RJ Jr: PCR amplification and restriction endonuclease analysis of a 65-kilodalton heat shock protein gene sequence for taxonomic separation of rapidly growing mycobacteria. *J Clin Microbiol* 1995, **33**:149–153.
23. Ringuet H, Akoua-Koffi C, Honore S, Varnerot A, Vincent V, Berche P, Gaillard JL, Pierre-Audigier C: *hsp65* Sequencing for Identification of Rapidly Growing Mycobacteria. *J Clin Microbiol* 1999, **37**:852–7.
24. Roth A, Fischer M, Hamid ME, Michalke S, Ludwig W, Mauch H: Differentiation of phylogenetically related slowly growing mycobacteria based on 16S-23S rRNA gene internal transcribed spacer sequences. *J Clin Microbiol* 1998, **36**:139–147.
25. Kim BJ, Lee SH, Lyu MA, Kim SJ, Bai GH, Chae GT, Kim EC, Cha CY, Kook YH: Identification on of Mycobacterial Species by Comparative Sequence Analysis of the RNA Polymerase Gene (*rpoB*). *J Clin Microbiol* 1999, **37**:1714–1720.
26. Domenech P, Jimenez MS, Menendez MC, Bull TJ, Samper S, Manrique A, Garcia MJ: *Mycobacterium mageritense* sp. nov. *Int J Syst Bacteriol* 1997, **47**:535–540.
27. Blackwood KS, He C, Gunton J, Turenne CY, Wolfe J, Kabani AM: Evaluation of *recA* sequences for identification of *Mycobacterium* species. *J Clin Microbiol* 2000, **38**:2846–2852.
28. Kasai H, Ezaki T, Harayama S: Differentiation of phylogenetically related slowly growing mycobacteria by their *gyrB* sequences. *J Clin Microbiol* 2000, **38**:301–308.
29. Zhang ZG, Ye ZQ, Yu L, Shi P: Phylogenomic reconstruction of lactic acid bacteria: an update. *BMC Evol Biol* 2011, **11**:1.
30. Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002, **30**:3059–3066.
31. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, **32**:1792–1797.
32. Wong YL, Ong CS, Ngeow YF: Molecular Typing of *Mycobacterium abscessus* Based on Tandem-Repeat Polymorphism. *J Clin Microbiol* 2012, **50**:3084–3088.
33. Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, **28**:27–30.
34. Rooney AP, Price NP, Ehrhardt C, Swezey JL, Bannan JD: Phylogeny and molecular taxonomy of the *Bacillus subtilis* species complex and description of *Bacillus subtilis* subsp. *inaquosorum* subsp. nov. *Int J Syst Evol Microbiol* 2009, **59**:2429–2436.
35. Schulenburg JH, Hurst GD, Huigens TM, van Meer MM, Jiggins FM, Majerus ME: Molecular evolution and phylogenetic utility of *Wolbachia* *ftsZ* and *wsp* gene sequences with special reference to the origin of male-killing. *Mol Biol Evol* 2000, **17**:584–600.
36. Casiraghi M, Bordenstein SR, Baldo L, Lo N, Beninati T, Wernegreen JJ, Werren JH, Bandi C: Phylogeny of *Wolbachia pipiensis* based on *gltA*, *groEL* and *ftsZ* gene sequences: clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the *Wolbachia* tree. *Microbiology* 2005, **151**:4015–4022.
37. Maddison WP: Gene trees in species trees. *Syst Biol* 1997, **46**:523–536.
38. Devulder G, Pérouse De Montclos M, Flandros JP: A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int J Syst Evol Microbiol* 2005, **55**:293–302.
39. Degnan JH, Rosenberg NA: Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* 2009, **24**:332–340.
40. Felsenstein J: *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates; 2005.
41. Delsuc F, Brinkmann H, Philippe H: Phylogenomics and the reconstruction of the tree of life. *Nature Rev Genet* 2005, **6**:361–375.
42. Bastian S, Veziris N, Roux AL, Brossier F, Gaillard JL, Jarlier V, Cambau E: Assessment of Clarithromycin Susceptibility in Strains Belonging to the *Mycobacterium abscessus* Group by *erm*(41) and *rml* Sequencing. *Antimicrob Agents Chemother* 2011, **55**:775.
43. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: Genbank. *Nucleic Acids Res* 2012, **40**:D48–D53.
44. Dai J, Chen Y, Lauzardo M: Web-Accessible Database of *hsp65* Sequences from *Mycobacterium* Reference Strains. *J Clin Microbiol* 2011, **49**:2296–2303.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, **215**:403–10.
46. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008, **9**:75.
47. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**:3389–3402.
48. Kim HY, Kook Y, Yun YJ, Park CG, Lee NY, Shim TS, Kim BJ, Kook YH: Proportions of *Mycobacterium massiliense* and *Mycobacterium bolletii* strains among Korean *Mycobacterium chelonae-Mycobacterium abscessus* group isolates. *J Clin Microbiol* 2008, **46**:3384–3390.
49. Little R, Dennis PP: Regulation of RNA Polymerase Synthesis Conditional Lethal Amber Mutations In The β Subunit Gene. *J Biol Chem* 1980, **255**:3536–3541.
50. Qamra R, Mandé SC, Coates ARM, Henderson B: The unusual chaperonins of *Mycobacterium tuberculosis*. *Tuberculosis* 2005, **85**:385–394.
51. Inaoka T, Matsumura Y, Tsuchido T: SodA and manganese are essential for resistance to oxidative stress in growing and sporulating cells of *Bacillus subtilis*. *J Bacteriol* 1999, **181**:1939–1943.
52. Walker GC: Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli*. *Microbiol Rev* 1984, **48**:60.
53. Limia A, Sangari FJ, Wagner D, Bermudez LE: Characterization and expression of *secA* in *Mycobacterium avium*. *Fems Microbiol Lett* 2001, **197**:151–157.
54. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentín F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, **23**:2947–2948.
55. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011, **28**:2731–2739.

56. Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Wörheide G, Baurain D: **Resolving difficult phylogenetic questions: why more sequences are not enough.** *PLoS Biol* 2011, **9**:e1000602.
57. Golubchik T, Wise MJ, Easteal S, Jermini LS: **Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments.** *Mol Biol Evol* 2007, **24**:2433–2442.
58. Landan G, Graur D: **Heads or tails: A simple reliability check for multiple sequence alignments.** *Mol Biol Evol* 2007, **24**:1380–1383.
59. R Core Team: *R: A language and environment for statistical computing.* Vienna: Austria: R Foundation for Statistical Computing; 2013.

doi:10.1186/1471-2164-14-879

Cite this article as: Tan *et al.*: A phylogenomic approach to bacterial subspecies classification: proof of concept in *Mycobacterium abscessus*. *BMC Genomics* 2013 **14**:879.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

