microbial biotechnology

Open Access

Combining comparative genomic analysis with machine learning reveals some promising diagnostic markers to identify five common pathogenic non-tuberculous mycobacteria

Xinmiao Jia,[1,2,†] Linfang Yang,[3,†] Cuidan Li,[4] Yingchun Xu,[2] Qiwen Yang[2,] and Fei Chen[4,5,6,7,*]**

[1] *Medical Research Center, State Key laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Peking Union Medical College, Beijing, 100730, China.*

[2] *Department of Clinical Laboratory, State Key laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, 100730, China.*

[3] *Departments of Dermatology, Affiliated Xingtai People's Hospital of Hebei Medical University, Xingtai, Hebei, 054001, China.*

[4] *CAS Key Laboratory of Genome Sciences & Information, China National Center for Bioinformation, Chinese Academy of Sciences, Beijing Institute of Genomics, Beijing, 100101, China.*

[5] *University of Chinese Academy of Sciences, Beijing, 100049, China.*

[6] *State Key Laboratory of Pathogenesis, Prevention, Treatment of High Incidence Diseases in Central Asia, Xinjiang, China.*

[7] *Beijing Key Laboratory of Genome and Precision Medicine Technologies, Beijing, China.*

Summary

Non-tuberculous mycobacteria (NTM) can cause various respiratory diseases and even death in severe cases, and its incidence has increased rapidly worldwide. To date, it's difficult to use routine diagnostic methods and strain identification to precisely diagnose various types of NTM infections. We combined systematic comparative genomics with machine learning to select new diagnostic markers for precisely identifying five common pathogenic NTMs (*Mycobacterium kansasii*, *Mycobacterium avium*, *Mycobacterium intracellular*, *Mycobacterium chelonae*, *Mycobacterium abscessus*). A panel including six genes and two SNPs (*nikA*, *benM*, *codA*, *pfkA2*, *mpr*, *yjcH*, *rrl* C2638T, *rrl* A1173G) was selected to simultaneously identify the five NTMs with high accuracy (> 90%). Notably, the panel only containing the six genes also showed a good classification effect (accuracy > 90%). Additionally, the two panels could precisely differentiate the five NTMs from *M. tuberculosis* (accuracy > 99%). We also revealed some new marker genes/SNPs/combinations to accurately discriminate any one of the five NTMs separately, which provided the possibility to diagnose one certain NTM infection precisely. Our research not only reveals novel promising diagnostic markers to promote the development of precision diagnosis in NTM infectious, but also provides an insight into precisely identifying various genetically close pathogens through comparative genomics and machine learning.

Introduction

Non-tuberculous mycobacteria (NTM) is a group of atypical mycobacteria other than *Mycobacterium tuberculosis* (Mtb) complex and *Mycobacterium leprae* (Johnson and Odell, 2014). They are widespread in the environment, and about one-third of them can infect people, mostly human lung tissue, leading to various respiratory diseases and even death in severe cases (Winthrop *et al.*, 2010; Johnson and Odell, 2014). In recent years, the incidence of NTM lung disease has increased rapidly worldwide (Reves and Schluger, 2014; Wu *et al.*, 2014)). From 2008 to 2015, the annual incidence of NTM lung disease increased from 3.13 to 4.73 per 100 000 persons, and the annual prevalence increased from 6.78 to 11.70 per 100 000 persons (Winthrop *et al.*, 2020). More seriously, many studies have shown that NTM could be transmitted from person to person, and its outbreaks have been reported in many hospitals (Aitken *et al.*, 2012; Bryant *et al.*, 2013).

Received 13 January, 2021; accepted 2 April, 2021.
For correspondence. *E-mail chenfei@big.ac.cn; Tel. +86 010 84097460; Fax: +86 010 84097720. **E-mail yangqiwen81@vip.163.com; Tel. +86 010 69159742; Fax: +86 010 69159766.
†These authors contributed equally to this study.
Microbial Biotechnology (2021) **14**(4), 1539–1549
doi:10.1111/1751-7915.13815

At present, it is difficult to use routine diagnostic methods (such as clinical symptoms, imaging characteristics, and biochemical indicators) to precisely diagnose various types of NTM clinical infections (Kwon and Koh, 2016). For one thing, the clinical symptoms, imaging characteristics, and biochemical indicators of most NTM infectious are usually similar to those of tuberculosis (Kwon and Koh, 2016). About 30% of NTM patients have been misdiagnosed and treated as multidrug-resistant tuberculosis at the beginning of treatment (Shahraki *et al*., 2015). For another thing, there are many types of pathogenic NTMs (*Mycobacterium avium*, *Mycobacterium intracellulare*, *Mycobacterium kansasii*, *Mycobacterium abscessus*, *Mycobacterium chelonei*, *etc*.), and related infectious diseases usually show similar clinical symptoms, imaging characteristics, and biochemical indicators (Griffith and Aksamit, 2016). These led to a high misdiagnosis rate for clinical NTM infections, further resulting in more serious clinical symptoms, prolonged course and even death (Gupta *et al*., 2020) since the treatments of TB and various NTM infections are different due to different drug-resistant spectrums (Gagneux, 2018).

On the other hand, accurate strain identification can promote the precision diagnosis of various types of NTM infections, but traditional strain identification technology is difficult to achieve the goal since only several NTM isolates can be identified through colonial morphology so far. Rapid-growing genetic testing has elevated the identification accuracy of various types of NTM isolates to some extent, with some common target genes such as 16S rRNA, *rpoB*, *hsp65* and *gyrA*/*gyrB* (Chimara *et al*., 2008; Unubol *et al*., 2019). However, due to high homology among various species of NTMs, the genetic testing results for different target genes are often inconsistent (Kim *et al*., 2018). Therefore, accurate strain identification of multiple types of NTMs and Mtb is still lacking.

Recent rapid development in next-generation sequencing technology, genomics, bioinformatics and big data analysis offers an opportunity to achieve the goal of accurate molecular typing of various microorganisms. To date, related studies about NTM only focussed on comparative genomic analyses within the same NTM species, which mainly revealed some genomic features (evolution, population structure, adaptation and virulence factors, *etc*.) in one certain NTM species like *Mycobacterium avium*/*Mycobacterium abscessus* (Sapriel *et al*., 2016; Yano *et al*., 2017). Thus, it is essential to search for the diagnostic markers to precisely identify various species of pathogenic NTM isolates through cross-species comparative genomic studies.

In this study, we conducted systematic comparative genomic and machine learning analyses (including pan-genome, random forest model and ensemble classification analyses) for five common pathogenic NTM species (*Mycobacterium kansasii* (Mka), *Mycobacterium avium* (Mav), *Mycobacterium intracellular* (Min), *Mycobacterium chelonae* (Mch), *Mycobacterium abscessus* (Mab)), which accounted for a considerable proportion (> 80%) of clinical NTM infections (Stacey et al., 2019). Pan-genome and comparative genomic analyses of 123 NTM complete genomes or assembled genomes at chromosome level (discovery set) were first performed to search for the specific core genes (SCGs) and specific core SNPs (SCSNPs) of the five NTMs. $\chi 2$ test and random forest model were then used to explore the marker genes/SNPs and the optimized combinations for discriminating any of the five NTMs through a larger validation set. Finally, an ensemble classification algorithm was performed to search for the panels that could simultaneously identify the five NTMs.

Results

Genomic features of five common pathogenic NTM species

One hundred and twenty-three complete genomes/assembled genomes at chromosome level from five common pathogenic NTM species were included in this study (7 Mka strains, 34 Mav strains, 30 Min strains, 8 Mch strains and 44 Mab strains) (Table S1). Among them, Mka, Mav and Min strains belong to slow-growing NTMs, in which Mav and Min are the members of Mycobacterium avium-intracellulare complex; Mch and Mab strains belong to rapid-growing NTM species and are the members of Mycobacterium chelonae-abscessus complex. Genome features are shown in Table 1. Compared with Mtb (~ 4.4 Mb; ~ 4100), the five NTMs possess larger genomic size (> 5 Mb) and more gene number (> 4700). Mka has the largest genome and the most genes (6.5 Mb; ~ 5800 genes), followed by Min (5.8 Mb; ~ 5400 genes). The other three NTM species contain a ~ 5 Mb genome and 4700–5000 genes. There is no obvious difference between rapid-growing and slow-growing NTM species in genome size and gene number. Besides, the GC content of three slow-growing NTM species (> 66%) is higher than that of Mtb (~ 65%), whereas the GC content of rapid-growing ones (~ 64%) is lower than that of Mtb.

We then performed phylogenetic analysis for the five NTMs (123) and Mtb strains (40) (Fig. 1). Figure 1 showed that different species of NTM isolates were clustered in different clades, which were obviously differentiated from the Mtb clade. Here the strains from M. avium-intracellulare complex (Mav and Min strains) and M. chelonae-abscessus complex (Mch and Mab strains) are clustered together compose one large clade differentiated from Mka.

Table 1. Genomic feature of the five common pathogenic NTM species.

		Strain number	Average genome size (Mb)	Average gene number (CDS)	Average GC content	Average core gene number	Average dispensable gene number	Average strain-specific gene number
Slow-growing NTM	*Mycobacterium kansasii* (Mka)	7	6.46	5756	66.01%	5234 (90.93%)	426 (7.40%)	96 (1.67%)
	Mycobacterium avium (Mav)	34	5.04	4704	69.18%	3786 (80.48%)	858 (18.24%)	60 (1.28%)
	Mycobacterium intracellulare (Min)	30	5.79	5406	67.82%	3881 (71.79%)	1402 (25.93%)	123 (2.28%)
Rapid-growing NTM	*Mycobacterium chelonei* (Mch)	8	5.04	4910	63.99%	3565 (72.61%)	1151 (23.44%)	194 (3.95%)
	Mycobacterium abscessus (Mab)	44	5.03	4948	64.16%	3699 (74.76%)	1170 (23.65%)	79 (1.60%)
Reference	*Mycobacterium tuberculosis* (Mtb)	40	4.41	4078	65.60%	3810 (93.43%)	263 (6.45%)	5 (0.12%)

These phylogenetic features agree with the results of comparative genomics analysis: The same species/complex/group of NTM isolates have higher ANIs and more homologous genes than different ones (Table S2A–C). The result of ANI analysis showed that the ANIs of the same NTM species are ~ 99%; the ANI between Mav and Min is ~ 86%; the ANI between Mab and Mch is ~ 83%. The ANIs between Mka and M. avium-intracellulare/M. chelonae-abscessus complexes are ~ 78% and ~ 70%, respectively; the ANIs between slow-growing and rapid-growing NTMs are 70–72% (Table S2A).

On the other hand, gene homology analysis (Table S2B, C) showed that the same species of NTM strains owned the highest homology (> 4000 homologous genes; > 89% of homologous gene ratios); same complexes of NTMs (M. avium-intracellulare/M. chelonae-abscessus) own higher homology (~ 4000 homologous genes, ~ 80% of homologous gene ratios) than different ones (~ 2000 homologous genes, ~ 40% of homologous gene ratios); same groups of NTMs (slow-/rapid-growing NTMs) possessed more homologous genes (> 3200) and higher homologous gene ratios (> 56%) than different ones (~ 2000 homologous genes, ~ 40% of homologous gene ratios). Incidentally, compared with Mtb, slow-growing NTMs (> 2500 homologous genes; > 62% of homologous gene ratios) showed higher homology than rapid-growing NTMs (~ 1700 homologous genes; ~ 43% of homologous gene ratios), which is consistent with the ANI analysis (Table S2A).

Pan-genome and comparative genomic analyses revealing specific core genes/SNPs (SCGs/SCSNPs) of the five common pathogenic NTM species

To search for SCGs and SCSNPs of the five NTMs, we adopted a pan-genome analysis strategy. Mtb strains were also analysed as controls. We first obtained the core genes of each NTM species by pan-genome analysis (Fig. S1). There are 6490 orthologous genes in the seven Mka strains including 5234 core genes (90.93%), 7925 orthologous genes in the 34 Mav strains including 3786 core genes (80.48%), 11 559 orthologous genes in the 30 Min strains including 3881 core genes (71.79%), 7733 orthologous genes in the eight Mch strains including 3565 core genes (72.61%), 12 472 orthologous genes in the 44 Mab strains including 3699 core genes (74.76%) and 4295 orthologous genes in the 40 Mtb strains including 3810 core genes (93.43%) (Table 1). In general, the percentage of core genes of NTM species is less than that of Mtb, while the percentage of strain-specific genes of NTM species is more than that of Mtb (Table 1, Figs S2, S3). This is mainly due to the conserved genome sequences (similarity >99%) of Mtb strains (Jia *et al.*, 2017).

We then conducted a comparative genomic analysis to look for the SCGs of five species of NTM isolates, which could discriminate any of the five NTMs and Mtb strains. The number of SCGs varies among the five NTM species: Mka strains possessed the most SCGs (1136), followed by Min (264), Mch (169), Mab (140) and Mav (92) (Fig. S1, Fig. 2).

On the other hand, we implemented the comparative genomic analysis to reveal the SCSNPs, which could discriminate one NTM species from the other four NTM species and Mtb. We detected the SCSNPs using 1514 core genes shared in all the 123 NTM and 40 Mtb strains (Fig. 2). Compared with reference genome (Mtb H37Rv: NC_000962), Mka possesses the most SCSNPs (55 858 SNPs on 913 genes), followed by Mav (20 860 SNPs on 679 genes), Min (16 256 SNPs on 590 genes), Mab (1550 SNPs on 71 genes) and Mch (1451 SNPs on 70 genes) (Fig. S1).

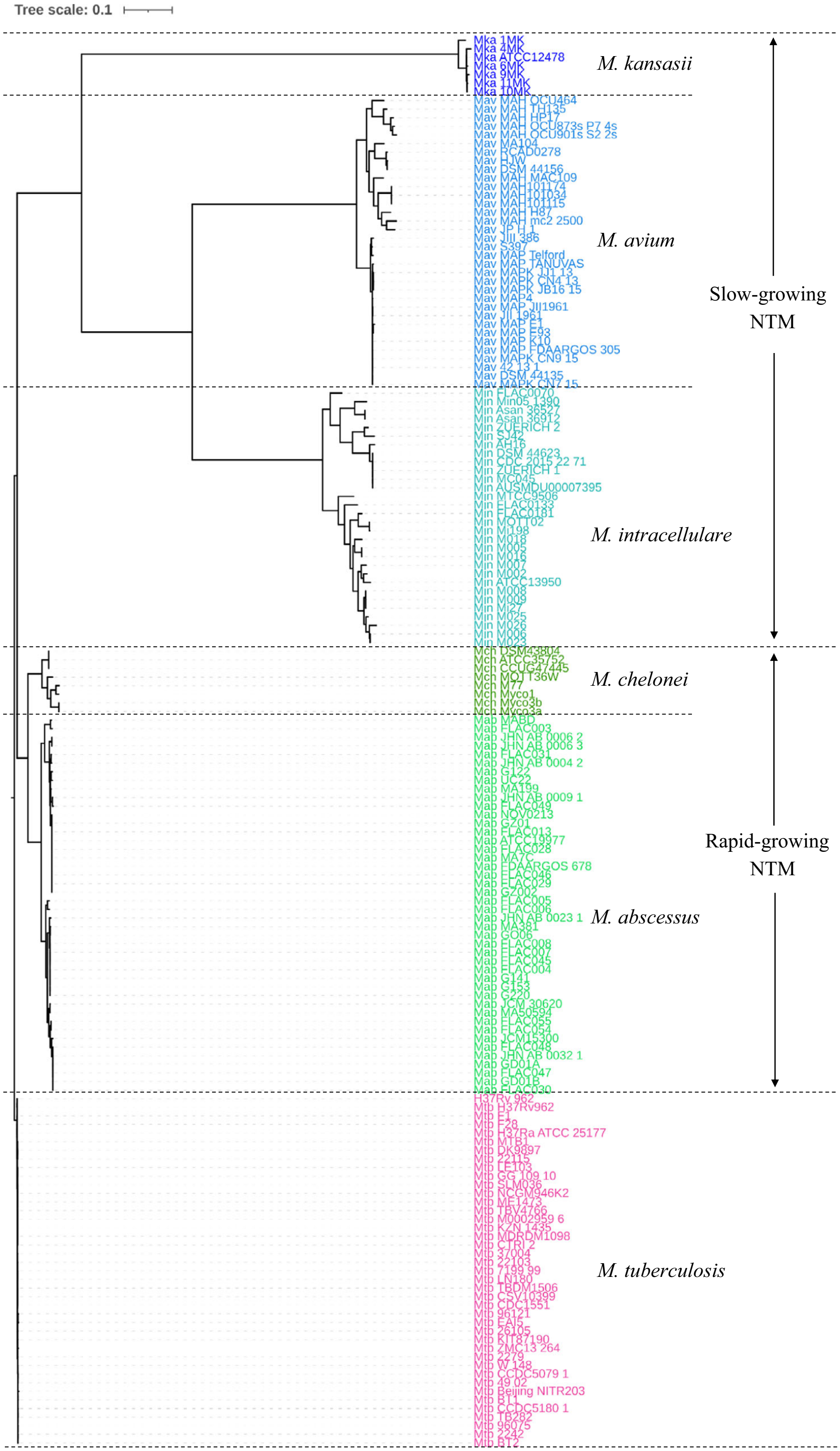


Fig. 1. Phylogenetic analysis of 123 NTM and 40 Mtb strains with complete genomes. The five common pathogenic NTMs are shown in different colours. The phylogenetic tree was constructed based on 289 751 core gene SNPs shared by these strains.

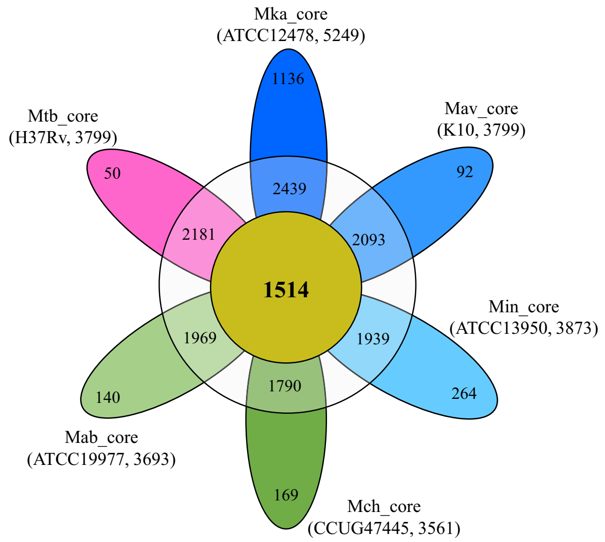


Fig. 2. Flower plot showing the core, dispensable, and species-specific genes of the five NTM species. Mtb were included as controls. The flower plot displays the core gene cluster number (in the centre), the dispensable gene number (in the annulus), and the species-specific gene number (in the petals) of the five NTM species. The numbers under the species name denote the core gene numbers of related species.

Although the number of SCGs and SCSNPs may be affected by the number of strains with complete genome sequences, they provide a reliable discovery set for subsequent screening of identification markers across NTM species by enlarging sample size. Incidentally, the more SCGs and SCSNPs in Mka species indicate the farther evolutionary distance from other NTM species (Fig. 1).

From discovery set to validation set: screening potential diagnostic markers to precisely identify any one of the five NTM species

To obtain a large validation set, we downloaded all the five NTM genome draft sequences from the NCBI SRA database (https://trace.ncbi.nlm.nih.gov/Traces/sra/), including 20 Mka, 159 Mav, 37 Min, 47 Mch and 285 Mab ones (Table S3). Genome sequences of 233 Mtb strains were also included as controls (Table S4). We first used $\chi 2$ tests to obtain marker SCGs and SCSNPs for identifying any of the five types of NTMs (p-value <0.01): 349 Mka, 91 Mav, 263 Min, 95 Mch and 139 Mab SCGs; 146 Mka, 305 Mav, 21 Min, 8 Mch and 28 Mab SCSNPs (Fig. S1).

Random forest algorithm was then used to explore the optimized gene/SNP combinations for identifying any of the five NTM species based on the above marker SCGs/SCSNPs. The results showed the optimized marker gene combinations to discriminate any of the five NTM species: two Mka marker genes (*pfkA2* and *MKAN_11495*), three Mav marker genes (*nikA*, *ddpC*, and *yejF*), three Min marker genes (*mnhF1*, *codA*, and *dmlR*), two Mch markers genes (*yjcH* and *mpr*) and five Mab marker genes (*benM*, *aqpZ*, *aldHT*, *osmX*, and *fsr*) (Fig. S1, Fig. 3, Table 2). The predictive ability of these gene combinations was further assessed using the area under the receiver operating characteristic curve (AUROC, AUC), which presented excellent predictive powers in the validating sets (AUC: 1.000 for Mka, 0.991 for Mav, 0.985 for Min, 0.955 for Mch and 0.952 for Mab; Sensitivity: 1.000 for Mka, 0.987 for Mav, 0.973 for Min, 0.957 for Mch and 0.909 for Mab; Specificity: 1.000 for Mka, 0.995 for Mav, 0.996 for Min, 0.952 for Mch and 0.996 for Mab) (Table 2). The predictive power of a single marker gene was then analysed, which also showed excellent predictive power (AUC > 0.95, Sensitivity > 0.9, Specificity > 0.95, Table S5). Overall, both single marker gene and marker gene combination could accurately identify any of the five species of NTM strains. In addition, these marker genes from the same combination showed similar AUC, sensitivity, and specificity values (Table S5).

By using random forest algorithm, we further screened out the optimized marker SNP combinations to discriminate the five NTM species, including eight Mka marker SNPs (*rrl* G377A, *rrl* C426T, *rrl* G2923A, *rrl* T3022C, *R1461* G2427C, *Rv2808* A28C, *Rv2808* A50C, *Rv2808* A111G), 11 Mav marker SNPs (*ino1* G832A, *clpB* G2124C, *rrl* G447A, *rrl* T455A, *rrl* G2368T, *rrl* T3066A, *Rv1461* G2133A, *acpM* C177T, *Rv2402* G570C, *rpsK* C90G, *pks13* G672C), six Min marker SNPs (*rpoC* T1282A, *rpoC* C1283G, *eccC5* G1908T, *acpM* C180G, *sdhA* C348G, *sdhA* G394T), two Mch markers SNPs (*rrl* C2638T, *rrl* G2654A) and two Mab marker SNPs (*rrl* A1173G, *aceE* C867T) (Fig. S1, Fig. 3, Table 2). The predictive abilities of the above marker SNP combinations (Table 2) were then analysed (AUC: 0.975 for Mka, 0.942 for Mav, 0.837 for Min, 0.756 for Mch and 0.922 for Mab; Sensitivity: 0.950 for Mka, 0.887 for Mav, 0.676 for Min, 0.553 for Mch and 0.846 for Mab; Specificity: 1.000 for Mka, 0.992 for Mav, 0.998 for Min, 0.958 for Mch and 0.996 for Mab; Fig. 3 and Table 2). We also analysed the predictive power of single marker SNP (Table S6), which was lower than the corresponding marker SNP combinations.

Importantly, we discovered that some marker SNPs on a 2000 bp region (1000–3000) of *rrl* gene coding for 23S ribosomal RNA could distinguish four species of NTMs (Mka, Mav, Mch, and Mab) with high accuracy (Table 2, Table S6), showing potential for accurately identifying some NTM species.

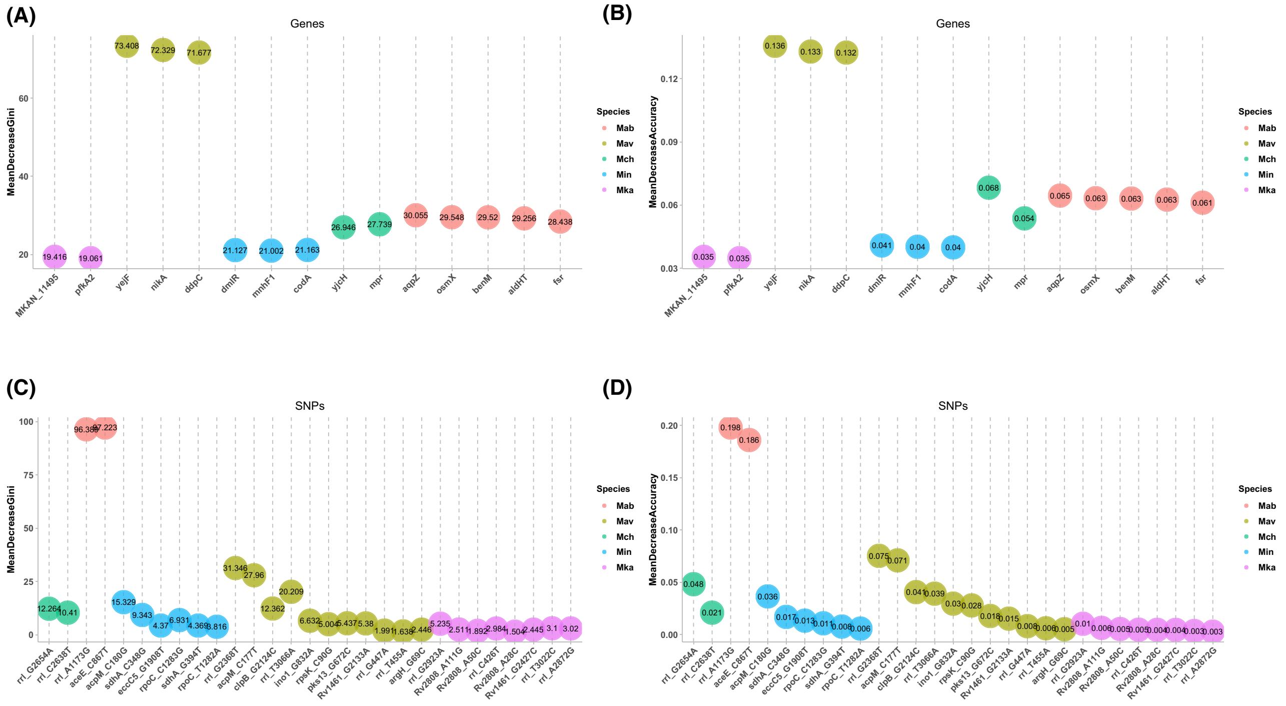


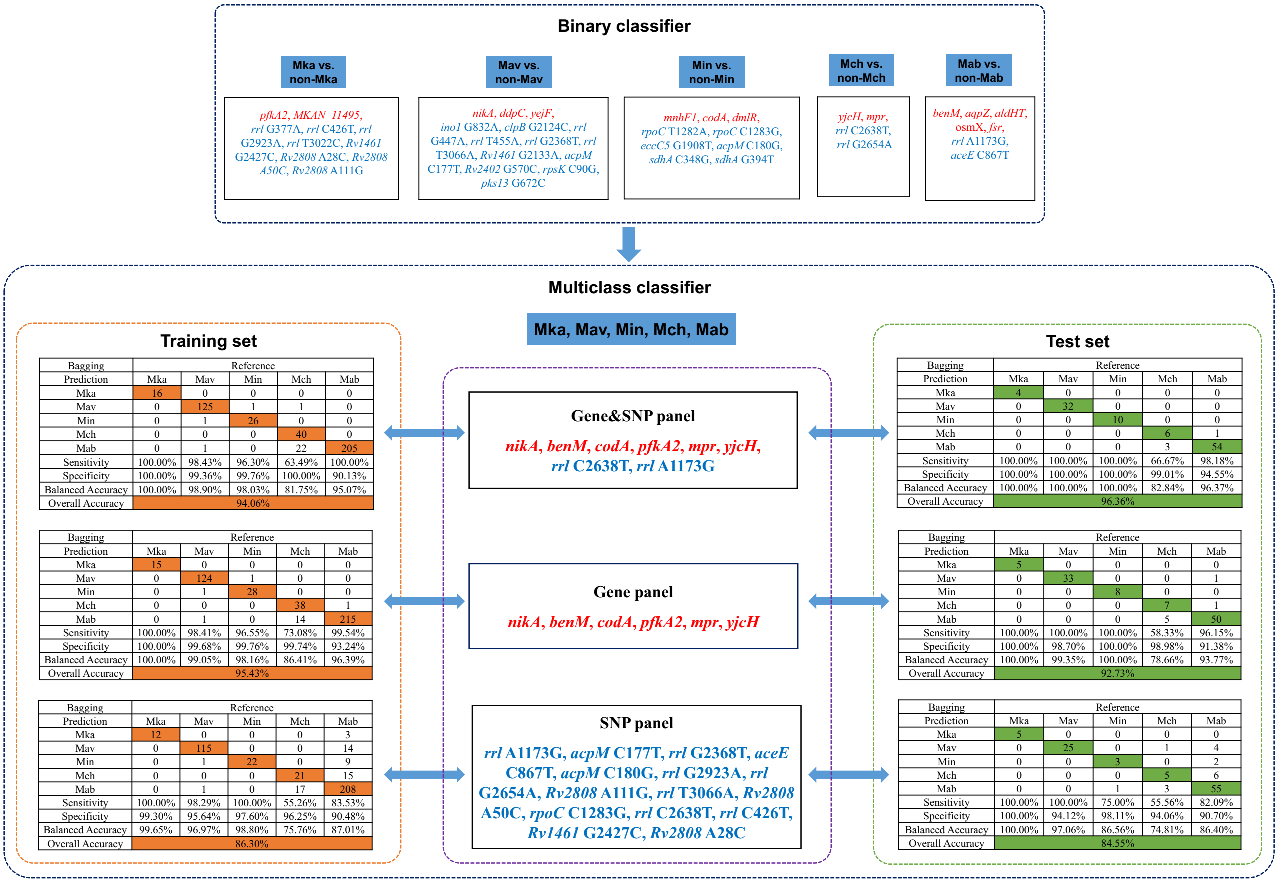
Fig. 3. Importance of the optimized gene/SNP combinations to identify the five common pathogenic NTM species using random forest models.
A. Mean Decrease Gini coefficient of the optimized gene combinations.
B. Mean Decrease Accuracy of the optimized gene combinations.
C. Mean Decrease Gini coefficient of the optimized SNP combinations.
D. Mean Decrease Accuracy of the optimized SNP combinations.

Table 2. Optimized gene/SNP combinations to identify the five common pathogenic NTM strains.

Species	Optimized gene/SNP combinations		Sensitivity (%)	Specificity (%)	AUC
Mka	Genes	*pfkA2*, *MKAN_11495*	1	1	1
	SNPs	*rrl* G377A, *rrl* C426T, *rrl* G2923A, *rrl* T3022C, *Rv1461* G2427C, *Rv2808* A28C, *Rv2808* A50C, *Rv2808* A111G	0.95	1	0.975
Mav	Genes	*nikA*, *ddpC*, *yejF*	0.987	0.995	0.991
	SNPs	*ino1* G832A, *clpB* G2124C, *rrl* G447A, *rrl* T455A, *rrl* G2368T, *rrl* T3066A, *Rv1461* G2133A, *acpM* C177T, *Rv2402* G570C, *rpsK* C90G, *pks13* G672C	0.887	0.992	0.942
Min	Genes	*mnhF1*, *codA*, *dmlR*	0.973	0.996	0.985
	SNPs	*rpoC* T1282A, *rpoC* C1283G, *eccC5* G1908T, *acpM* C180G, *sdhA* C348G, *sdhA* G394T	0.676	0.998	0.837
Mch	Genes	*yjcH*, *mpr*	0.957	0.952	0.955
	SNPs	*rrl* C2638T, *rrl* G2654A	0.553	0.958	0.756
Mab	Genes	*benM*, *aqpZ*, *aldHT*, *osmX*, *fsr*	0.909	0.996	0.952
	SNPs	*rrl* A1173G, *aceE* C867T	0.846	0.996	0.922

Ensemble classification analysis revealed gene/SNP panels to identify the five common pathogenic NTM species simultaneously

To further achieve good classification performance for the five NTM species simultaneously, an ensemble classification algorithm was adopted based on the preselected genes and SNPs in each RF binary classifier (Fig. 3, Table 2, Fig. S1). A gene&SNP panel, a gene panel, and an SNP panel were screened out for simultaneously identifying the five NTM species with high accuracy (Fig. 4). The gene&SNP panel (*nikA*, *benM*, *codA*, *pfkA2*, *mpr*, *yjcH*, *rrl* C2638T and *rrl* A1173G; overall accuracy > 94%) and the gene panel (*nikA*, *benM*, *codA*, *pfkA2*, *mpr* and *yjcH*; overall accuracy > 92%) own higher classification accuracy than the SNP panel (*rrl* A1173G, *acpM* C177T, *rrl* G2368T, *aceE* C867T, *acpM* C180G, *rrl* G2923A, *rrl* G2654A, *Rv2808* A111G, *rrl* T3066A, *Rv2808* A50C, *rpoC* C1283G, *rrl* C2638T, *rrl* C426T, *Rv1461* G2427C, and *Rv2808*



Binary classifier

Mka vs. non-Mka

pfkA2, *MKAN_11495*, *rrl* G377A, *rrl* C426T, *rrl* G2923A, *rrl* T3022C, *Rv1461* G2427C, *Rv2808* A28C, *Rv2808* *A50C*, *Rv2808* A111G

Mav vs. non-Mav

nikA, *ddpC*, *yejF*, *ino1* G832A, *clpB* G2124C, *rrl* G447A, *rrl* T455A, *rrl* G2368T, *rrl* T3066A, *Rv1461* G2133A, *acpM* C177T, *Rv2402* G570C, *rpsK* C90G, *pks13* G672C

Min vs. non-Min

mnhF1, *codA*, *dmlR*, *rpoC* T1282A, *rpoC* C1283G, *eccC5* G1908T, *acpM* C180G, *sdhA* C348G, *sdhA* G394T

Mch vs. non-Mch

yjcH, *mpr*, *rrl* C2638T, *rrl* G2654A

Mab vs. non-Mab

benM, *aqpZ*, *aldHT*, osmX, *fsr*, *rrl* A1173G, *aceE* C867T

Multiclass classifier

Mka, Mav, Min, Mch, Mab

Training set

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	16	0	0	0	0
Mav	0	125	1	1	0
Min	0	1	26	0	0
Mch	0	0	0	40	0
Mab	0	1	0	22	205
Sensitivity	100.00%	98.43%	96.30%	63.49%	100.00%
Specificity	100.00%	99.36%	99.76%	100.00%	90.13%
Balanced Accuracy	100.00%	98.90%	98.03%	81.75%	95.07%
Overall Accuracy	94.06%				

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	15	0	0	0	0
Mav	0	124	1	0	0
Min	0	1	28	0	0
Mch	0	0	0	38	1
Mab	0	1	0	14	215
Sensitivity	100.00%	98.41%	96.55%	73.08%	99.54%
Specificity	100.00%	99.68%	99.76%	99.74%	93.24%
Balanced Accuracy	100.00%	99.05%	98.16%	86.41%	96.39%
Overall Accuracy	95.43%				

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	12	0	0	0	3
Mav	0	115	0	0	14
Min	0	1	22	0	9
Mch	0	0	0	21	15
Mab	0	1	0	17	208
Sensitivity	100.00%	98.29%	100.00%	55.26%	83.53%
Specificity	99.30%	95.64%	97.60%	96.25%	90.48%
Balanced Accuracy	99.65%	96.97%	98.80%	75.76%	87.01%
Overall Accuracy	86.30%				

Gene&SNP panel

nikA, *benM*, *codA*, *pfkA2*, *mpr*, *yjcH*, *rrl* C2638T, *rrl* A1173G

Gene panel

nikA, *benM*, *codA*, *pfkA2*, *mpr*, *yjcH*

SNP panel

rrl A1173G, *acpM* C177T, *rrl* G2368T, *aceE* C867T, *acpM* C180G, *rrl* G2923A, *rrl* G2654A, *Rv2808* A111G, *rrl* T3066A, *Rv2808* A50C, *rpoC* C1283G, *rrl* C2638T, *rrl* C426T, *Rv1461* G2427C, *Rv2808* A28C

Test set

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	4	0	0	0	0
Mav	0	32	0	0	0
Min	0	0	10	0	0
Mch	0	0	0	6	1
Mab	0	0	0	3	54
Sensitivity	100.00%	100.00%	100.00%	66.67%	98.18%
Specificity	100.00%	100.00%	100.00%	99.01%	94.55%
Balanced Accuracy	100.00%	100.00%	100.00%	82.84%	96.37%
Overall Accuracy	96.36%				

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	5	0	0	0	0
Mav	0	33	0	0	1
Min	0	0	8	0	0
Mch	0	0	0	7	1
Mab	0	0	0	5	50
Sensitivity	100.00%	100.00%	100.00%	58.33%	96.15%
Specificity	100.00%	98.70%	100.00%	98.98%	91.38%
Balanced Accuracy	100.00%	99.35%	100.00%	78.66%	93.77%
Overall Accuracy	92.73%				

Bagging	Reference				
Prediction	Mka	Mav	Min	Mch	Mab
Mka	5	0	0	0	0
Mav	0	25	0	1	4
Min	0	0	3	0	2
Mch	0	0	0	5	6
Mab	0	0	1	3	55
Sensitivity	100.00%	100.00%	75.00%	55.56%	82.09%
Specificity	100.00%	94.12%	98.11%	94.06%	90.70%
Balanced Accuracy	100.00%	97.06%	86.56%	74.81%	86.40%
Overall Accuracy	84.55%				

Fig. 4. Ensemble classification workflow for data generation and analysis of Gene/SNP panels for simultaneously discriminating the five common pathogenic NTM species. The multiclass classifier was proposed based on the above RF binary classifier ('non-Mka' indicates the Mav, Min, Mch, Mab and Mtb strains; 'non-Mav' indicates the Mka, Min, Mch, Mab and Mtb strains; 'non-Min' indicates the Mka, Mav, Mch, Mab and Mtb strains; 'non-Mch' indicates the Mka, Mav, Min, Mab and Mtb strains; 'non-Mab' indicates the Mka, Mav, Min, Mch and Mtb strains). Confusion matrixes of the gene/SNP panels in the training set and test set were shown on the left and right, respectively.

A28C; overall accuracy > 84%). Here the three panels show higher classification accuracy for Mka, Mav and Min, followed by Mab; whereas they have the lowest classification accuracy for Mch (Fig. 4).

Discussion

In this study, we combined systematic comparative genomics with machine learning (including pan-genome, random forest model and ensemble classification) to screen out some diagnostic markers for precisely identifying five common pathogenic NTM species (Mka, Mav, Min, Mch and Mab). A panel including six genes and two SNPs (*nikA*, *benM*, *codA*, *pfkA2*, *mpr*, *yjcH*, *rrl* C2638T and *rrl* A1173G) was selected to simultaneously identify the five NTM species with high accuracy (> 95% for Mka, Mav, Min and Mab; > 80% for Mch; Fig. 4). The panel only containing the six genes also showed a good classification effect on the five NTM species (Fig. 4). The two panels could also precisely differentiate the five NTMs from Mtb (accuracy > 99%). Overall, the two panels provide novel promising diagnostic markers to promote the development of precision diagnosis in human infectious diseases caused by NTM. Here, we noticed the lower prediction accuracy for Mch in the two panels (Fig. 4), which might be due to fewer complete genomes of Mch strains in NCBI. Larger sample size can optimize this in the future. In addition, our studies also revealed 15 new marker genes, 29 new marker SNPs, and ten optimized combinations to accurately discriminate any one of the five NTM species separately (Fig. 3, Table 2, Tables S5, S6), which provided the possibility of how to diagnose one certain NTM infection precisely.

In general, our study showed a lower prediction accuracy of marker SNPs than that of marker genes (Table 2), indicating the better NTM strain identification effect by marker genes. Functional analysis revealed that many of these marker genes were related to transmembrane transporter activity (*nikA*, *ddpC*, *mnhF1*, *yjcH* and *osmX*), rapid growth (*aqpZ*) and drug resistance (*fsr*) (https://www.uniprot.org/). They might be responsible for host/environmental/antibiotic adaptation of NTM strains (https://www.uniprot.org/) and further partially contributed to the individuality of different types of NTM strains. As a result, they showed better classification performance on various kinds of NTMs than marker SNPs.

Importantly, our research shows the potential of machine learning on the detection and identification of biomarkers from big data. Machine learning has been applied to precision medicine progressively because it can improve the prediction accuracy of patterns/features by automated training for algorithms that learn from genomic data (Goecks *et al.*, 2020). In this study, we first revealed some SCGs and SCSNPs by comparative genomic analysis, and two important algorithms of machine learning (random forest and ensemble classification) were further adopted to screen out the species-biomarkers in a large validation set. First, to obtain the optimized marker genes/SNPs to identify any of the five NTM species, a binary classifier belonging to machine learning, random forest algorithm was selected to discriminate one NTM species from others (Mka vs. 'non-Mka', Mav vs. 'non-Mav', Min vs. 'non-Min', Mch vs. 'non-Mch', and Mab vs. 'non-Mab'). Random forest algorithm is a flexible, easy-to-use machine learning algorithm that produces, even without hyper-parameter tuning, a good result of binary classifier (Blanchet *et al.*, 2020). To further achieve good classification performance for the five NTM species simultaneously, an ensemble classification algorithm was adopted based on the pre-selected genes and SNPs in each binary classifier. Here, ensemble classification is a powerful machine learning tool capable of achieving the excellent performance of multiple classifiers, which can correct for errors made by any individual classifier and lead to better accuracy overall (Bramer, 2013).

In summary, our research not only reveals some new panels and marker genes/SNPs to accurately discriminate the five NTM species, but it also provides an insight into precisely identifying various genetically close species of pathogens through comparative genomics and machine learning. Indeed, these panels and markers warrant further confirmation and optimization with larger sample size studies.

Experimental procedures

Bacterial genomes

To completely understand the genomic features of different NTM species, we collected the ones with more than five completed genomes or assembled genome sequences at the chromosome level in NCBI. The complete genome sequences and assembled genome sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) up to March 12, 2021. Here, one complete genome (*M. kansasii* Kuro-I) was deleted due to the suspicious gene number (8110), which is much more than that of other *M. kansasii* strains (~ 5700). In total, one complete genome and six assembled genomes at chromosome level of seven *Mycobacterium kansasii* (Mka) strains, 33 complete genomes and one assembled genome at chromosome level of 34 M. *avium* (Mav) strains, 30 complete genomes of 30 M. *intracellular* (Min) strains, 5 complete genomes and three assembled genomes at chromosome level of 8 M. *chelonae* (Mch) strains, and 40 complete genomes and four assembled genomes at chromosome level of 44 M. abscessus (Mab) strains were obtained (Table S1). The raw Illumina sequencing reads of genomic DNAs of NTM species were also downloaded from the SRA database as a validation set. In total, 20 Mka strains, 159 Mav strains, 37 Min strains, 47 Mch strains and 285 Mab strains with a total base of more than 400 Mb were downloaded (Table S3). For comparison, complete genomes of 40 randomly selected *M. tuberculosis* (Mtb) strains (Table S1) were also analysed as a control in the discovery data set, and complete genomes of the other 233 Mtb strains were included in the validation data set (Table S4).

Genome re-annotation, Average nucleotide identity (ANI) and gene homology analysis

All the downloaded genome sequences were re-annotated with Prokka (Seemann, 2014). The protein functions were further annotated using Blast2GO (https://www.blast2go.com/). Pairwise ANI was calculated using pyani 0.2.10 with ANIb (Pritchard *et al.*, 2016). Gene homology analysis was analysed using BLAT with a threshold of identity 50% and coverage 50%, and Inparanoid/multiparanoid (Remm *et al.*, 2001).

Phylogenetic analysis

The phylogenetic analysis of the NTM and Mtb strains was based on the core gene SNPs detected by MUMmer 3.23 (Delcher *et al.*, 2002)) using *M. tuberculosis* H37Rv (NC_000962) as the reference. The MAFFT (Nakamura *et al.*, 2018) was adopted to align the concatenated SNP sequences and phylogenetic tree was generated by FastTree (Price *et al.*, 2009).

Identification of Mka/Mav/Min/Mch/Mab specific core genes (SCGs)

All proteins from the 7 Mka, 34 Mav, 30 Min, 8 Mch, 44 Mab and 40 Mtb strains were clustered by the pan-genome pipeline Roary to create a multiFASTA alignment of core genes using PRANK and a fast core gene alignment with MAFFT, and paralogs are not split (Andrew *et al*., 2015), respectively. The characteristic curves of NTM/Mtb pan-genome, core-genome and new genes were depicted by Pan-Genome Profile Analyze Tool (PanGP) (Zhao *et al*., 2014) with DG-sampling algorithms. Gene homology analysis was further conducted among core genes of different types NTM species (Mka/Mav/Min/Mch/Mab/Mtb) using Inparanoid/multiparanoid (Remm *et al*., 2001). The SCGs of Mka were defined as those present in all Mka strains but absent from all other strains containing Mtb. The Mav/Min/Mch/Mab SCGs were identified using similar parameters.

Identification of Mka/Mav/Min/Mch/Mab specific core SNPs (SCSNPs)

SNPs were detected by MUMmer 3.23 (Delcher *et al*., 2002) using *Mycobacterium tuberculosis* H37Rv (NC_000962) as the reference. The Mka SCSNPs were defined as those present in all Mka strains but absent from all other strains containing Mtb. The Mav SCSNPs were defined as those present in all Mav strains but absent from all other strains containing Mtb. The Min/Mch/Mab SCSNPs were identified using similar parameters. SNPs were further annotated according to the *.gff file generated by Prokka (Seemann, 2014).

Validation of Mka/Mav/Min/Mch/Mab SCGs and SCSNPs

To verify the SCGs obtained above, the raw Illumina sequencing reads downloaded from the SRA database were mapped to the sequences of SCGs of each NTMs using BWA 0.5.9 (Li and Durbin, 2009), respectively. Only SCGs with depth more than 10X and coverage of more than 80% were considered to exist in the strain. To verify the SCSNPs, the raw Illumina sequencing reads were mapped to the genome sequences of Mtb H37Rv (NC_000962) using BWA 0.5.9 ((Li and Durbin, 2009), respectively. SNPs were analysed using SAMtools 0.1.19 (Li *et al*., 2009) and VarScan (Koboldt *et al*., 2009), and filtered with at least ten reads covered and 70% supported.

To verify the SCGs/SCSNPs of Mka, all downloaded data were classified into two groups (Mka and non-Mka). Categorical data were expressed as 0 or 1, categorical variables using χ^2 tests. All analyses were performed using R software, and differences were considered statistically significant at $P < 0.01$. The Mav/Min/Mch/Mab SCGs were screened using a similar procedure.

Markers identification and panel screening using random forest and ensemble classification algorithm of machine learning

To build strain identification models of each kind of NTM, a random-forest classification algorithm of machine learning was further conducted using the screened results obtained from $\chi2$ tests. This analysis was performed using the randomForest package in R and the number of trees grown was 10 000. 70% and 30% of samples were randomly selected as training group and testing group, respectively. Cross-validation was conducted using rfcv () function with parameter 'cv.fold=10' and 'step=0.8'. Important features were extracted using the imp () function based on the value of Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). AUCs were calculated by receiver operating characteristic (ROC) analysis using the roc () function of pROC package in R. To simultaneously achieve good classification performance on five NTM groups, an ensemble classification algorithm of machine learning, 'bagging ()' from the package 'adabag' in R, was proposed based on the above RF binary classifier (4/5 for training and 1/5 for testing) using pre-selected features in each binary classifier obtained from random forest. The analysis process is shown in Fig. S1.

Funding Information

This work was supported by Special Projects of Infectious Diseases for National Key Research and Development Program of China (2018ZX10302301-004-003); the National Natural Science Foundation of China (32000463); the Research Fund of Peking Union Medical College Hospital (pumch201912054).

Conflict of interest

The authors declare that they are inventors on various patent applications covering several of the methods and results reported here.

Author contributions

XJ, QY and FC designed the study. XJ, LY, and CL performed bioinformatics analyses. XJ, LY, YX, QY and FC prepared the manuscript. All authors contributed to and approved the final manuscript.

References

Andrew, J.P., Carla, A.C., Martin, H., Vanessa, K.W., Sandra, R., and Matthew, T.G.H. *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31:** 3691–3693.

Aitken, M.L., Limaye, A., Pottinger, P., Whimbey, E., Goss, C.H., Tonelli, M.R., *et al.* (2012) Respiratory outbreak of *Mycobacterium abscessus* subspecies massiliense in a lung transplant and cystic fibrosis center. *Am J Respir Crit Care Med* **185:** 231–232.

Blanchet, L., Vitale, R., van Vorstenbosch, R., Stavropoulos, G., Pender, J., Jonkers, D., *et al.* (2020) Constructing bi-plots for random forest. *Tutorial. Anal Chim Acta* **1131:** 146–155.

Bramer, M. (2013) Ensemble classification. *Principles of Data Mining. Undergraduate Topics in Computer Science*. London: Springer, pp. 209–220.

Bryant, J.M., Grogono, D.M., Greaves, D., Foweraker, J., Roddick, I., Inns, T., *et al.* (2013) Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet* **381:** 1551–1560.

Chimara, E., Ferrazoli, L., Ueky, S.Y., Martins, M.C., Durham, A.M., Arbeit, R.D., and Leao, S.C. (2008) Reliable identification of mycobacterial species by PCR-restriction enzyme analysis (PRA)-hsp65 in a reference laboratory and elaboration of a sequence-based extended algorithm of PRA-hsp65 patterns. *BMC Microbiol* **8:** 48.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30:** 2478–2483.

Gagneux, S. (2018) Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* **16:** 202–213.

Goecks, J., Jalili, V., Heiser, L.M., and Gray, J.W. (2020) How machine learning will transform biomedicine. *Cell* **181:** 92–101.

Griffith, D.E., and Aksamit, T.R. (2016) Understanding nontuberculous mycobacterial lung disease: it’s been a long time coming. *F1000Research* **5:** 2797.

Gupta, N., Mittal, A., Muhammed Niyas, V.K., Banerjee, S., Ray, Y., Kodan, P., *et al.* (2020) Nontuberculous mycobacteria: a report of eighteen cases from a tertiary care center in India. *Lung India* **37:** 495–500.

Jia, X., Yang, L.i., Dong, M., Chen, S., Lv, L., Cao, D., *et al.* (2017) The bioinformatics analysis of comparative genomics of *Mycobacterium tuberculosis* Complex (MTBC) provides insight into dissimilarities between intraspecific groups differing in host association, virulence, and epitope diversity. *Front Cell Infect Microbiol* **7:** 88.

Johnson, M.M., and Odell, J.A. (2014) Nontuberculous mycobacterial pulmonary infections. *J Thorac Dis* **6:** 210–220.

Kim, J.U., Ryu, D.S., Cha, C.H., and Park, S.H. (2018) Paradigm for diagnosing mycobacterial disease: direct detection and differentiation of *Mycobacterium tuberculosis* complex and non-tuberculous mycobacteria in clinical specimens using multiplex real-time PCR. *J Clin Pathol* **71:** 774–780.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25:** 2283–2285.

Kwon, Y.S., and Koh, W.J. (2016) Diagnosis and treatment of nontuberculous mycobacterial lung disease. *J Korean Med Sci* **31:** 649–659.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34:** 2490–2492.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26:** 1641–1650.

Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* **8:** 12–24.

Remm, M., Storm, C.E.V., and Sonnhammer, E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041–1052.

Reves, R., and Schluger, N.W. (2014) Update in tuberculosis and nontuberculous mycobacterial infections 2013. *Am J Respir Crit Care Med* **189:** 894–898.

Sapriel, G., Konjek, J., Orgeur, M., Bouri, L., Frézal, L., Roux, A.-L., *et al.* (2016) Genome-wide mosaicism within *Mycobacterium abscessus*: evolutionary and epidemiological implications. *BMC Genom* **17:** 118.

Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30:** 2068–2069.

Shahraki, A.H., Heidarieh, P., Bostanabad, S.Z., Khosravi, A.D., Hashemzadeh, M., Khandan, S., *et al.* (2015) "Multidrug-resistant tuberculosis" may be nontuberculous mycobacteria. *Eur J Intern Med* **26:** 279–284.

Stacey, L., Martiniano, M. D., Jerry, A., Nick, M. D., Charles, L., and Daley, M. D. (2019) *Nontuberculosis mycobacterial disease. Kendig’s Disorders of the Respiratory Tract in Children* (9th edn). Robert W. W., Robin D., Albert L., Felix R., Peter S., Heather J. Z., Andrew B. (eds). Elsevier, pp. 498–506.

Unubol, N., Kizilkaya, I.T., Okullu, S.O., Koksalan, K., and Kocagoz, T. (2019) Simple identification of mycobacterial species by sequence-specific multiple polymerase chain reactions. *Curr Microbiol* **76:** 791–798.

Winthrop, K.L., Marras, T.K., Adjemian, J., Zhang, H., Wang, P., and Zhang, Q. (2020) Incidence and prevalence of nontuberculous mycobacterial lung disease in a large U.S. Managed Care Health Plan, 2008–2015. *Ann Am Thorac Soc* **17:** 178–185.

Winthrop, K.L., McNelley, E., Kendall, B., Marshall-Olson, A., Morris, C., Cassidy, M., *et al.* (2010) Pulmonary nontuberculous mycobacterial disease prevalence and clinical features: an emerging public health disease. *Am J Respir Crit Care Med* **182:** 977–982.

Wu, J., Zhang, Y., Li, J., Lin, S., Wang, L., Jiang, Y., *et al.* (2014) Increase in nontuberculous mycobacteria isolated in Shanghai, China: results from a population-based study. *PLoS One* **9:** e109736.

Yano, H., Iwamoto, T., Nishiuchi, Y., Nakajima, C., Starkova, D.A., Mokrousov, I., *et al.* (2017) Population structure and local adaptation of mac lung disease agent *Mycobacterium avium* subsp. hominissuis. *Genome Biol Evol* **9:** 2403–2417.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., *et al.* (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **30:** 1297–1299.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. The screening process of some potential diagnostic markers (SNP/gene) for the five common pathogenic NTM species. One hundred and twenty-three NTM complete genomes from NCBI were first performed the comparative genomic analysis to obtain the discovery set of SCGs/SCSNPs for the five common NTM species. 548 NTM strains with raw genome sequencing reads and 233 Mtb strains were further analysed as the validation set to screen some potential diagnostic markers (SNPs/genes).

Fig. S2. Gene accumulation curves of the pan-genome (blue) and core-genome (green) of five common pathogenic NTMs and Mtb. The blue boxes denote the pan-genome size for each genome for comparison. The green boxes show the core-genome size for each genome for comparison. The curve is the least-squares fit of the power law for the average values.

Fig. S3. Curve (red) for the number of new genes with an increase in the number of five common pathogenic NTMs and Mtb.

Table S1. Information of NTM strains with complete genome sequences used in this study.

Table S2. (A) Pairwise comparison of ANIs among the five NTMs. (B) Pairwise comparison of homologous gene numbers among the five NTMs. (C) Pairwise comparison of homologous gene ratio among the five NTMs.

Table S3. Information of NTM strains downloaded from SRA used for further validation.

Table S4. Information of Mtb strains with complete genome sequences used for further validation.

Table S5. Single-gene markers from optimized combinations to identify the five common pathogenic NTM strains.

Table S6. Single-SNP markers from optimized combinations to identify the five common pathogenic NTM strains.