**ABSTRACT**

*Mycobacterium tuberculosis* (MTB) is mostly found in humans, and it can cause more than two million deaths each year with increasing morbidity. Although lineages of MTB show identical nucleotide relationships, they have different characteristics such as evolution, transmission, drug resistance, host interaction, latency, and vaccine effectiveness. It is necessary to have better understanding of MTB relationships based on similarities in genome sizes and phylogenetic analysis. This paper observes the relationships of MTB based on nucleotide through phylogenetic frameworks. The MTB species consist of six lineages, and each lineage has various size of genomes. This difference contributes to virulence of MTB affecting levels of severity, morbidity, and mortality of diseases. Genetic diversity of MTB can contribute to global threats in the world such as outbreak of tuberculosis, Multi Drug Resistant (MDR) and Extensively Drug Resistant (XDR) tuberculosis.

**INTRODUCTION**

Tuberculosis, an infectious disease for humans caused by *Mycobacterium tuberculosis* (MTB) bacillus is a global health problem with high prevalence. The 2018 Global Tuberculosis Report found 9.6 million new cases of the tuberculosis in the world, estimated to be 254 cases per 100,000 populations. In Indonesia,
tuberculosis cases in 2017 were 420,994 cases. Approximately, 2.5% of tuberculosis patients had died each year, and more than 480,000 incidences of MDR TB and 1.7 billion latent infections were found.\textsuperscript{1,2} Currently, Indonesia is on the third rank of six countries that have the most tuberculosis cases in the world and become High Burden Countries (HBC) country based on three indicators: TB, TB/HIV, and Multi Drug-Resistant Tuberculosis (MDR-TB).\textsuperscript{1}

Genotypes of MTB are responsible for clinical phenotype findings and determinants of MTB virulence factors. Strains of different MTB species show different cellular and clinical phenotypes. For example, strains of lineage 5 and lineage 6 are metabolically slower-growing and less virulent, while lineage 2 and lineage 4 are more virulent in terms of disease severity and have fast human-to-human transmission. Besides, different genomic loci will be associated with appearance of phenotypic differences in cellular and clinical terms.\textsuperscript{3,4} Adequate TB control can be conducted by understanding characteristics of populations and host genetics. However, it is also necessary to understand the genetic (phylogenetic) diversity of pathogens.\textsuperscript{4,5}

Phylogenetics describes an evolution of a living being related to its morphology, physiology, and character development. By phylogenetics, levels of the kinship of a living being can be determined based on its mutations in expression, gene duplication, as well as classification of incomplete lineages.\textsuperscript{5} There are several methods of genomic determination based on nucleotide sequences (DNA) such as whole genome sequencing (WGS), IS6110 DNA fingerprinting with restriction fragment length polymorphism (RFLP) analysis, spoligotyping, mycobacterial interspersed repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing, genomic deletion analysis with regions of differences (RDs), multi locus sequence analysis (MLSA), large sequence polymorphisms (LSPs) and single nucleotide polymorphisms (SNPs).\textsuperscript{3,7,8} Any changes in the MTB genome will affect its virulence. This microorganism comprises of 4.4 million base pairs encoding 4,000 genes.\textsuperscript{9}

Based on WGS analysis along the MTB genome, 1,200 SNPs, clustered regularly interspaced short palindromic repeats (CRISPR) with highly polymorphic features, VNTR, and insertion sequences (IS) are found. A broad range of the PE/PPE gene group will impact expansions of the ESAT-6 gene cluster and affect the virulence of MTB.\textsuperscript{11}

Analysis of the relationship between bacterial species was conducted by using a theoretical approach by determining levels of certain bacterial groups based on their hierarchy. There was no specific agreement used to determine the hierarchy in MTB, so it was difficult to find an exact relationship between genotype and phenotype and between strains of these species. Levels of microorganisms in MTB also did not have a standard nomenclature for grouping bacteria. The only way to determine the relationship of any MTB strain was to use the WGS, clearly defining the MTB phylogeny. Therefore, in this paper, in the discussion, the phylogeny frameworks with WGS analysis becomes a basis for lineage stratification for other phylogeny frameworks with different analytical methods.\textsuperscript{12}

**MTB phylogeny based on whole genome sequencing (WGS)**

This particular sequencing method identifies, measures and compares all pictures of genes such as DNA sequences, structural variations, gene expressions, or annotations of regulatory and functional elements on genomic scales.\textsuperscript{13} Phylogeny analysis compiles epidemiological studies by using multilocus analysis of nucleotide MTB sequences by CRISPR, VNTR (spoligotyping), and MIRU-VNTR techniques.\textsuperscript{14} Principles of the CRISPR technique is to repeat sequences of a spacer, and these repetitions must accord to length and sequences of DNA. Sometimes a difference in the repetitions can be found, but these are very rare.\textsuperscript{15} The VNTR technique can find mutations in a DNA sequence repeatedly. Therefore, this technique can be used to identify and discriminate against a kinship of living beings.\textsuperscript{16} Gene examination with this technique is conducted precisely and sensitively.
because of stability in target gene locus. WGS method in the phylogenetic analysis uses 108 MTB strains spread throughout the world. This phylogeny has similarities with strains analysed by the LSP method. The MIRU-VNTR and LSP analysis shows a significant close level of kinship, and they are considered to have a solid basis for the classification of phylogeny by other methods. 

Comas et al. compiled the MTB phylogeny framework that infects humans into six lineages (1 to 6). In the phylogeny framework, the MTB is divided into several lineages and strains based on the spoligotyping analysis. The East-African-Indian EAI (EAI) strain is a strain from lineage 1, and other strains such as Beijing family and Central-Asian (CAS) are strains from lineage 2 and lineage 3 respectively. Cameroon, Uganda, X, Haarlem, and Latin-American-Mediterranean (LAM) strains are a group of strains from lineage 4. Meanwhile, tAFR1 and 2 strains (known as Mycobacterium africanum strains) are strains from lineage 5 and 6.

The WGS phylogeny framework forms a basis of the phylogeny analysis of Sreevatsan et al., which uses the basis of polymorphism analysis of the two coding genes as he classified MTB ancestry into three groups of principal genetic group (PGG). MTB in lineages 1, 2, 3, 5, and 6 are organisms that enter PGG1. MTB strains in lineage 4 are organisms that enter PGG2, and strains H37Rv, T16, T78, T38, and T60 are groups of strains that enter PGG3.

**Phylogeny MTB based on MLSA**

Analyses of relationships of each MTB strain in several places have important values for local public health and form a basis of clinical epidemiological research. In molecular epidemiology research and analyses of transmission of the strains, DNA sequencing of MTB plays important roles. MLSA, a strain identification technique, applies a procedure by tagging for a protein-coding gene for basic function of MTB. During the process, Mycobacterium experiences only a few purification selections. Therefore, it is often found that there is a genetic shift which impacts on diversity of functions of the *Mycobacterium*. The presence of genetic diversity in the MTB population followed by an increase of human populations, urbanizations, and cross-country travels can be a cause for emergence and spread of drug-resistant TB.

The MLSA-based phylogeny framework is constructed based on an analysis of data from a large set of coding gene sequences collected globally. Based on this analysis, the phylogeny framework is divided into lineages of Philippines, Indian Ocean, West Africa 1 and 2, India and East Africa, East Asia, Europe, and the Americas. Each group of strains is labelled according to its dominance in a specific geographic area. For example, M. Canetti has many differences with MTB (showing a truncated branching line), so there is a separation of group between the MTB and M. Canetti. Ancient strain groups such as Philippines, India Ocean, and West Africa are spread by sea, and modern strain groups such as India and East Africa, East Asia, Europe, and the Americas are spread by land. Although the Mycobacterium strains that infect animals are ecotypically different, these strains may represent a portion of genetic diversity found in all MTB infecting humans. Therefore, this strain are also included in the MTB phylogeny framework.

**MTB phylogeny based on large sequence polymorphism (LSP)**

LSP analysis is an analysis of gene identification based on findings of insertion or deletion of different genes along the MTB genomes. The phylogeny analysis compiles genomic deletion in 875 strains from 80 countries. Essence of this phylogeny analysis obtains six main lineages from the MTB, namely East Asian, East-African-Indian, Euro-American, Indo-Oceanic, West African-1. Then West African 2 lineages come from a common ancestor and have a history of evolution in the humans.

The lineages in this phylogeny belong to the same group of strains as those reported by previous studies. Indo-oceanic, an ancient
strain group for a region of the TbD1 gene, is not found in modern strain groups such as East-Asia (including the Beijing strain family), West-African 1 and 2 belong to the Mycobacterium africanum strain group, while Euro-America is included in the PGG 2 and 3 strain groups according to the classification of Sreevatsan et al. 20

Comas et al. conducted a study evaluating human movements in Africa, and it suggested that genetic population of MTB was associated with certain geographic structures. Therefore, names of each lineage also reflect this geographic relationship. For example, lineage 1 is very common in East African countries, the Philippines, and the Indian Ocean. Meanwhile, lineages 2, 3, 4, and 5 are widely available around East Asia, East Africa- Central Asia, European-American, and West Africa respectively. 24

**MTB phylogeny based on single nucleotide polymorphisms (SNP)**

Occurrences of genetic variations within a population can cause polymorphisms related to differences in genetic traits or phenotypes due to a stimulus from the environment. 25 Single nucleotide polymorphisms (SNP) are genetic variations often found in a population. Each SNP represents a difference in one nucleotide (for example; a replacement of cytosine nucleotide with thymine nucleotide in one DNA sequence). 26 This phylogeny framework analyses 36 SNPs from 5069 MTB isolates in four different populations such as New York, New Jersey, Houston, and Finland. There are nine clusters with one new cluster called II. A (located between clusters II and III). This cluster classification has a relationship with the three PGG classifications of Sreevatsan et al 20, namely PGG 1 organisms included in clusters I and II, PGG2 organisms included in clusters III and VI, and PGG3 organisms included in clusters VII and VIII. 24

**MTB phylogeny based on spoligotyping**

Spoligotyping is the first PCR-based genotyping method used in a large-scale MTB complex (MTBC). The target method is a direct repetition locus found in almost all MTBC strains. These loci are members of a large cluster of family of bacteria with regularly palindromic short-sequence repeating loci. Each locus contains a distinct sequence of spacer sequences, varying between strains and separated by repeating motifs (called direct repeats). The spacer set on the isolates is amplified with primary pair as opposed to the direct repeats flanking each spacer. The amplifiers obtained are then hybridized (in a reference set of 43 spacers from M. tuberculosis H37Rv and M. bovis BCG) and placed on a membrane spoligotyping kit (to indicate presence or absence of any reference spacers of the strains tested). However, a classic spoligotyping requires high skills at the hybridization stage although it is relatively inexpensive. 27

MTB phylogeny classification based on spoligotyping describes the MTB relationship by spoligotype analysis of nine different clades. The clade presents all lineages of the organism and all its offspring starting from its ancestor to all its derivative species. 28 The spoligotype pattern shows that more than 13,008 isolates are grouped into 813 same types (90% intact isolates). There are six main rules of classification (A to F) to obtain a better phylogeny organization. Based on the six classification rules, there are a formation of 36 main clades of MTB isolates and nine groups of strains. Some main identified clades are the Beijing clade, the EAI clade, the Haarlem clade, the LAM clade, CAS, the European clade (with low IS6110 markers), the X clade (mostly found in the US and UK), the T clade (marked by the absence of spacers 33-36). The nine strains of this classification are Mycobacterium africanum, Beijing, M. bovis, EAI, CAS, T, Haarlem, X, and LAM (96.9% of similarity based on the phylogenetic relationship of the MTB family). 29

The Beijing strain is one of causes of tuberculosis drug resistance in the world. Increasing prevalence of this strain has become an important issue in controlling the tuberculosis. Interestingly, this spoligotype clade, the Mycobacterium africanum strain, has a high
proportion of 6% of all types of *Mycobacterium tuberculosis* in Africa.\[^{20,29}\]

**MTB phylogeny based on PGG**

The PGG method classifies isolates into one of three groups based on variants in katG and gyrA genes that are not identical. Two MTB-coding genes, namely codon katG 463 and gyrA codon 95 show high polymorphism without involvement of anti-tuberculosis drug resistance.\[^{20}\] Based on the polymorphism in these two genes, Sreevatsan et al. conducted a strain analysis of 6000 isolates in Houston and New York.\[^{20}\] The results found 48 large clusters which were grouped based on their genotypes into PGP 1, 2, and 3. Observations on PGG showed that each organism experienced a decrease in transmissibility and virulence. PGG1 group organisms, a group of strains with evolutionary histories and characteristics similar to M. bovis, are a cause of Bovine Tuberculosis. This includes special strains of M. bovis isolate, New York city IS6110 type W strain, Houston IS6110 type 002, 003, 007, 015, and 003. Some strains included in PGG2 are Erdman strain, New York city strain C, Houston IS6110 type 004, 006, 016, 020, and 030. Meanwhile, other MTB strains are comprised in PGG3 are H37Ra, H37Rv strains, and Houston IS6110 type 001.20 strains.

Several points of polymorphism in the genes are generally used for an analysis of phylogeny studies. However, this is rarely conducted in drug resistance analysis. Polymorphisms commonly occur with genetic mutations related to metabolism and drug resistance. MTB from the Beijing strain is a main strain in the world with the highest incidence of drug resistance of any other MTB strains.\[^{30,31}\] All MTBC members are assigned to one of three distinctive PGG groups based on the location of the polymorphisms occurring at the two marked gene sites. The polymorphisms of the M. bovis, M. microti, and M. africanum isolates that were analysed had characteristics according to PGG1.\[^{20}\] The following table shows a summary of MTB relationships with different analyses (Table 1).

<table>
<thead>
<tr>
<th>PHYLOGENY</th>
<th>WGS[^{14}]</th>
<th>MLSA[^{22}]</th>
<th>LSP[^{18}]</th>
<th>SNP[^{27}]</th>
<th>Spoligotyping[^{29}]</th>
<th>PGG[^{20}]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lineage 2</td>
<td>East Asia</td>
<td>East-Asian</td>
<td>Cluster II</td>
<td>Beijing, ST523, ST623</td>
<td>PGG1</td>
<td></td>
</tr>
<tr>
<td>Lineage 3</td>
<td>India and East Africa</td>
<td>East-African-Indian</td>
<td>Cluster II A</td>
<td>CAS</td>
<td>PGG1</td>
<td></td>
</tr>
<tr>
<td>Lineage 4</td>
<td>Europe and Americas</td>
<td>Euro-American</td>
<td>Cluster III and VII</td>
<td>X, Haarlem, LAM, Uganda</td>
<td>PGG2-PGG3</td>
<td></td>
</tr>
<tr>
<td>Lineage 1</td>
<td>Rim of Indian Ocean - the Philippines</td>
<td>Indo-Oceanic</td>
<td>Cluster I</td>
<td>EAI</td>
<td>PGG1</td>
<td></td>
</tr>
<tr>
<td>Lineage 5</td>
<td>West Africa</td>
<td>M. africanum West-African1</td>
<td>-</td>
<td>AFR1</td>
<td>PGG1</td>
<td></td>
</tr>
<tr>
<td>Lineage 6</td>
<td>West Africa</td>
<td>M. africanum West-African2</td>
<td>-</td>
<td>AFR2</td>
<td>PGG1</td>
<td></td>
</tr>
</tbody>
</table>

**CONCLUSION**

The phylogenetic framework portrays the genotypic and phenotypic relationships among the MTB lineages and sublineages scattered around the world, arranged based on differences in genome sizes, polymorphisms and geographic location. This kinship greatly influences the virulence of MTB and determines interactions between the host and its pathogens. Of the various scattered MTB species, six MTB lineages...
are arranged in a phylogenetic framework by using kinship analysis methods such as WGS, MLSA, LSP, SN, Spoligotyping, PGG.

CONFLICT OF INTEREST
No conflict of interest.

ACKNOWLEDGEMENT
None declare.

REFERENCES
10. Roychowdhury T, Mandal S, Bhattacharya A. Analysis of IS6110 insertion sites provide a glimpse into genome evolution of Mycobacterium tuberculosis. Scientific Reports. 2015;5:12567.
17. Jagielski T, Van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current methods in the molecular typing of mycobacterium tu-