A Genomic Signature for Genotyping *Mycobacterium tuberculosis*

David Tarazona, Luis Jaramillo, Victor Borda, Kelly Levano, Marco Galarza, Heinner Guio*

1Laboratorio de Biotecnología y Biología Molecular, Centro Nacional de Salud Pública, Instituto Nacional de Salud, Lima, Perú. Heinner Guio - E-mail: heinnerguio@gmail.com; *Corresponding author

Received July 7, 2017; Revised July 13, 2017; Accepted July 23, 2017; Published July 31, 2017

Abstract: *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis (TB), has a vast diversity of genotypes including Beijing, CAS, EAI, Haarlem, LAM, X, Ural, T, AFRI1 and AFRI2. However, genotyping can be expensive, time consuming and in some cases, results may vary depending on methodology used. Here, we proposed a new set of 10 SNPs using a total of 249 MTB genomes, and selected by first the inclusion/ exclusion (IE) criteria using spoligotyping and phylogenies, followed by the selection of the nonsynonymous SNPs present in the most conserved cluster of orthologous groups (COG) of each genotype of *MTB*. Genotype assignment of the new set of 10 SNPs was validated using an additional of 34 MTB genomes and results showed 100% correlation with their known genotypes. Our set of 10 SNPs have not been previously reported and cover the *MTB* genotypes that are prevalent worldwide. This set of SNPs could be used for molecular epidemiology with drug resistant markers.

Keywords: Genomic signature, Genotyping, *Mycobacterium tuberculosis*

Background: Tuberculosis (TB), responsible for approximately 1.4 million deaths annually, represents one of the main challenges for public health. The decoding of the *M. tuberculosis* (MTB) genome [1] has accelerated the advances in its genetic diversity and TB diagnosis. *MTB* families like Beijing, Latin American (LAM), Haarlem, Central Asian (CAS), T, East-African-Indian (EAI), Euro-American (X-type) and AFRI have been established via MIRU-VNTR and spoligotyping [2-5]. These techniques are based on the number of tandem intergenic repeats in the genome of *MTB* [6] and the amplification by PCR of clustered regulatory short palindromic repeats (CRISPR) [7], respectively. The development of genotyping methodologies has allowed: (1) comparison of strains circulating in different populations, (2) characterization of outbreaks, (3) distinction between recent and old transmissions, (4) detection of cross-contamination in laboratories, (5) detection of re-infection or relapse and (6) identification of populations with high risk of transmission allowing subsequent implementation of appropriate control strategy [8-10]. At present, there are regions with higher frequency of TB infection and a particular *MTB* genotype [11]. However, countries with a heterogeneous population, such as the United States [12], Peru [13], Israel [14], European countries [15, 16] and even China [17], have reported a greater diversity of *MTB* families which complicate their distribution and evolution. This is where genotyping becomes a challenge with the current methodologies MIRU-VNTR and spoligotyping, which lack sufficient discriminatory power to differentiate between families [18]. Whole genome sequencing (WGS) has been shown to give superior resolution to that of MIRU-VNTR and spoligotyping providing all possible genomic targets, information on drug resistance, genome evolution and virulence determinants [19]. However, massive sequencing costs are still expensive in countries with high TB incidence. The best approach will be to analyze a specific set of SNPs associated to a determined genotype. Recent studies have proposed sets of SNPs, one using 45 SNPs from 04 genomic sequences [20] and another using 62 SNPs from 1601 genomes [21]. However, these two studies have been based on phylogenies. In this study, we proposed a new set of 10 SNPs using a total of 249 *MTB* genomes selected first by the inclusion/ exclusion (IE) criteria using spoligotyping and phylogenies, followed by the selection of the nonsynonymous SNPs present in the most conserved COG (cluster of orthologous groups) of each genotype (Beijing, CAS, EAI, Haarlem, LAM, X, Ural, T, AFRI1 and AFRI2) of *MTB*. The addition of spoligotyping provides higher informative results on the phylogeographic distribution of *MTB*‘s genotypic diversity [22].
Methodology:

**Genome sequences selection:**
A total of 249 MTB genomic sequences were used in this study: 125 of the genomic sequences had known genotypes determined by the gold standard method (spoligotyping), including Beijing (lineage 2), CAS (lineage 3), EAI (lineage 1 y 3), Haarlem (lineage 4), LAM (lineage 4), X genotypes (lineage 4), Ural (lineage 4), T (lineage 4), AFR11 (lineage 5) and AFR12 (lineage 6), based on publications and public databases; 76 of the genomic sequences had known genotypes determined by phylogenetic analysis including lineage 1 (EAI), lineage 2 (Beijing), lineage 3 (CAS and EAI), lineage 4 (Haarlem, LAM, T, X), lineage 5 (AFR11) and lineage 6 (AFR12); and 48 of the genomic sequences has unknown genotypes. These genomic sequences were obtained from public domains of the international database from the National Institute of Health (NIH) ([ftp://ftp.ncbi.nlm.nih.gov/genomes/](http://www.ncbi.nlm.nih.gov/genomes/)), European Nucleotide Archive (ENA; [http://www.ebi.ac.uk/ena/](http://www.ebi.ac.uk/ena/)) and Kyoto Encyclopedia of Genes and Genomes (KEGG) ([http://www.genome.jp/en/gn_ftp.html](http://www.genome.jp/en/gn_ftp.html)). The MTB sequences analyzed were from 35 countries (USA, Canada, China, Colombia, El Salvador, Ethiopia, Gambia, Germany, Ghana, Guatemala, India, Italy, Japan, Kazakhstan, Malaysia, Mexico, Mongolia, Nepal, Netherlands, Nicaragua, Panama, Peru, Philippines, Puerto Rico, Russia, Sierra Leone, South Africa, Sweden, Tanzania, Iran, Thailand, Uganda, United Kingdom, Uruguay and Vietnam) allowing us to obtain a global diversity and MTB representativeness. It is a retrospective study so ethical approval was not required.

Additionally, another 34 MTB genomic sequences with known genotypes: Beijing (n=12), LAM (n=07), X (n=06), CAS (n=04), Haarlem (n=04) and EAI (n=01) were used for genotype confirmation using our new proposed set of 10 SNPs. Analyzed data were mapped to the H37Rv reference genome (number accession: NC_000962.3) using BWA ([23](http://bioinformatics.oop/bioinformatics-OOP.html)). SAMtools ([24](http://bioinformatics.oop/bioinformatics-OOP.html)), kSNP ([25](http://bioinformatics.oop/bioinformatics-OOP.html)) and ParSNP were used to identify SNPs. These programs exclude the analysis of SNPs in non-sequenced regions.

**Phylogenetic analysis of genomic sequences:**
Phylogenetic assays were performed on 173 MTB genomes (125 with genotypes determined by spoligotyping and 48 with unknown genotypes) using two different programs of alignment-free: kSNP ([26](http://bioinformatics.oop/bioinformatics-OOP.html)) and ParSNP ([27](http://bioinformatics.oop/bioinformatics-OOP.html)). Each program uses different algorithms to select SNPs, constructs the Maximum likelihood phylogenies and compares topologies. kSNP identifies SNPs based on k-mers analysis without using a reference genome. This program requires a k-value for the mer size, which is calculated by Kchooser tool. Also, in order to count the k-mers in the genomes, kSNP uses the jellyfish software ([28](http://bioinformatics.oop/bioinformatics-OOP.html)), then compares these mers in all the genomes to find SNPs and finally create a SNP matrix to make phylogenetic trees. ParSNP uses maximal unique matches (MUM) to generate a multiple genome alignment and a SNP tree. This MUM for two genomes is a genetic index that considers the level of DNA conservation of the core genome and the proportion of DNA shared by these genomes. After the alignment process, the core-genome SNPs are selected and used to construct a phylogenetic tree.

**Cog analysis:**
To determine the most variable and conserved COGs for MTB, a COG analysis was performed on 249 MTB genomes that included the genotypes Beijing, CAS, EAI, Haarlem, LAM, X, Ural, AFR11, AFR12 and unknowns. Briefly, a comparative analysis based on H37Rv to determine that SNP was performed using SamTools program ([24](http://bioinformatics.oop/bioinformatics-OOP.html)). Then, a SNP database was generated which included SNPs corresponding gene and COG.

**SNP set selection:**
Genotyping assignment was performed in two steps: (I) using the inclusion/exclusion criteria in experimental genotype (Spoligotypes) and phylogenetic prediction and, (II) using COG.

(I) **Inclusion/ exclusion (IE) criteria using spoligotyping and phylogenies:**
First, 125 MTB genomes with experimental information of genotypes were processed. Each genome sequence was transformed to binary codes, as presence (1) or absence (0) of a mutation in loci, based on H37Rv. Then, genomes were grouped according to their genotypes (Beijing, CAS, EAI, Haarlem, LAM, X, AFR11 and AFR12). Using IE criteria, loci shared with more than one genotype were discarded. Second, to reinforce the SNP selection, 124 additional MTB genomes (76 with genotypes determined by phylogenetic analysis and 48 of unknown genotypes with genotypes predicted by phylogenetic analysis) were included. Finally, the SNPs shared in more than one MTB genotype of Beijing, CAS, EAI, Haarlem, LAM, X, AFR11 and AFR12 were discarded. A total of 2123 specific SNPs to each genotype of MTB were identified as SNPs-IE.

(II) **Nonsynonymous SNPs selection based on COG (cluster of orthologous groups):**
Then, the SNPs were organized in COG functional groups ([29](http://bioinformatics.oop/bioinformatics-OOP.html)). The genotypes of the MTB genomes have shown variant locus distribution in COG starting from the most variable to the most conserved (S, R, I, Q, E, C, F, L, N, G, H, K, P, J, T, O, F, D, V, U and A). In this case, we selected the locus from the most conserved to the most variable gene family. Then, we selected a locus associated with a nonsynonymous SNP for each of the 10 MTB genotypes.

**Results & Discussion:**
**Lineage determination of 48 unknown MTB genomes by phylogenetic analysis:**
Tree topology was performed on 249 MTB genomes using kSNP and Parsnp ([Figure 1](http://bioinformatics.oop/bioinformatics-OOP.html)). Eight clades were identified with good support: AFR11, EAI, CAS, Beijing, Haarlem, LAM and T. However, clade AFR12 remains paraphyletic with strains at different levels of the bifurcation. Of the 48 unknown MTB genomes, 03 were AFR11, 16 were EAI, 03 were CAS, 22 were Beijing, 15 were Haarlem, 03 were X and 15 were LAM. One Haarlem strain (Haarlem/NITR2) as determined by spoligotyping fell in the X clade by both methodologies (kSNP and Parsnp). There are limitations in the assignment of genotypes by phylogenetic analysis and often require global information of the genomic sequences for an optimal lineage approximation.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 13(7): 224-230 (2017)
Identification and selection of SNP set:  
To develop the assigning system of SNPs and genotypes, we first used the IE criteria. 125 MTB genomes with known genotypes determined by spoligotyping were analyzed as described in Methods. The T genotype was not analyzed because the H37Rv falls in this genotype and was used in the process of genomic mapping. Additionally, 124 MTB genomes were included with genotypes determined by phylogenetic analysis. In a database, we integrated the 7649 SNPs from 249 genomic sequences of MTB that included the genotypes Beijing, CAS, EAI, Haarlem, LAM, X, Ural, AFRI1 and AFRI2. The number of non-specific genotype SNPs were 458, 854, 674, 624, 424, 412, 886, 1735 and 1582 respectively. After applying the IE criteria, the SNPs were organized in Group A, which contained SNPs unique for each genotype, and in Group B, which contained the SNPs shared between the different genotypes. Then, it was followed by a manual exclusion of shared SNPs. The result was a new group of SNPs (SNPs-IE) for the genotypes Beijing (n=20), CAS (n=276), EAI (n=208), Haarlem (n=15), LAM (n=37), X (n=16), Ural (n=372), AFRI1 (n=580) and AFRI2 (n=599) (Figure 2).

Finally, this was followed by the selection of the nonsynonymous SNPs present in the most conserved COG of each of MTB. Before the analysis, the COGs most variables and most conserved for each genotype of MTB were determined (Figure 3). Then, a SNP database was generated, which included SNPs, corresponding gene and COG, showing 2, 36, 20, 4, 2, 2, 40, 65 and 67 SNPs for Beijing, CAS, EAI, Haarlem, LAM, X, Ural, AFRI1 and AFRI2 respectively. We selected 10 nonsynonymous SNPs for the 9 MTB genotypes analyzed (Table 1 and 2) and an alternative SNPs set (Table 2).
Figure 2. Diagram of signature SNPs for MTB genotypes: Beijing, CAS, EAI, Haarlem, LAM, T, X, Ural, AFRI1 and AFRI2 after IE criteria.

Figure 3. The relative distribution of SNPs of Mycobacterium tuberculosis in protein belonged to certain Clusters of Orthologous Groups (COGs). Function unknown (S), General function prediction only (R), Lipid transport and metabolism (I), Secondary metabolites biosynthesis, transport, and catabolism (Q), Amino acid transport and metabolism (E), Energy production and conversion (C), Cell wall/membrane/envelope biogenesis (M), Replication, recombination and repair (L), Cell motility (N), Carbohydrate transport and metabolism (G), Coenzyme transport and metabolism (H), Inorganic ion transport and metabolism (P), Transcription (K), Translation, ribosomal structure and biogenesis (J), Post-translational modification, protein turnover, and chaperones (O), Signal transduction mechanisms (T), Nucleotide transport and metabolism (F), Cell cycle control, cell division, chromosome partitioning (D), Defense mechanisms (V), Intracellular trafficking, secretion, and vesicular transport (U), RNA processing and modification (A).
Validating genotype assignment using new proposed set of 10 SNPs:
Using an additional 34 MTB genomes with known genotypes Beijing (n=12), LAM (n=07), X (n=06), CAS (n=04), Haarlem (n=04) and EAI (n=01) we tested the ability of our new set of 10 SNPs to assign genotypes. There was 100% correlation of genotypes assignment with all the strains tested (Table 3).

Table 1: Stepwise SNP set selection: IE, differential genotype ; *, the loci belonging to other strains from different genotypes and uncommon between genotypes were eliminated. **, SNP set based in COG group (A→K); *** SNP set based in less variable COG group genotype.

| STEPS | Genotype | Beijing | CAS | EAI | Haarlem | LAM | X | Ural | AFR1 | AFR1 | T | Total |
|-------|----------|---------|-----|-----|---------|-----|----|------|------|------|---|-------|
| (1)   | Integration of 7649 SNPs | MTB genome sequences | 52 | 06 | 23 | 19 | 78 | 09 | 01 | 06 | 05 | 249 |
| (2)   | IE criteria selection | SNP set based in COG group** | 69 | 310 | 220 | 33 | 93 | 56 | 396 | 737 | 621 | 2535 |
| (3)   | Nonsynonymous SNPs selection present in most conserved COG | SNP set based in MTB most conservative | 20 | 276 | 15 | 16 | 37 | 572 | 580 | 599 | 2123 |
|       | Proposed SNP set *** | 01 | 01 | 01 | 01 | 01 | 01 | 01 | 02 | 02 | 10 |

Table 2. Set of 10 SNPs proposed to genotype MTB.

| Genotype | COG | Gene (locus tag) | SNP proposed |
|----------|-----|------------------|--------------|
| Beijing  | CAS | GuaB2 (Rv3411c)  | 3830349 (Ala391Thr: GGC→AGC) |
|          |     | (Rv3921c)        | 4409954 (Ala393Gly:GCC→GCG) |
|          | EAI | SecE1 (Rv0638)   | 734116 (Met127Thr: ATG→ACC) |
| Haarlem  | LAM | Hpt (Rv3625c)    | 4063682 (Leu191Met: CTG→ATG) |
|          |     | Smc (Rv2922c)    | 3236230 (Arg236Leu:CGT→CTT) |
|          | X   | Transcription    | 2946570 (Gly63Asp:GAC→GCG) |
|          |     | (Rv2618)         | 4063682 (Leu191Met:CTG→ATG) |
|          | Ural| RNA processing and modification | 1225462 (Asp228Gly:GAC→GCG) |
|          | AFR1| RNA processing and modification | 3858894 (Leu266Phe:CTT→TTT) |
|          | AFR2| RNA processing and modification | 1226021 (Gln42Asp:GAC→GCG) |
|          | Beijing| Post-translational modification, protein turnover, and chaperones | 4130604 (Ser42Asn:AGC→GCG) |
|          |     | IrtA (Rv1348)    | 1651308 (Glu198Gly: GAA→GCA) |
|          | CAS | IrtA (Rv1348)    | 1513189 (Ala48Val: GCA→GTA) |
|          | EAI | Defense mechanisms | 1956930 (Thr39Pro: ACT→CTT) |
|          |     | (Rv1730)         | 1513189 (Ala48Val: GCA→GTA) |
|          | Haarlem| Transcription | 765150 (Gly594Glu: GGG→GGG) |
|          |     | RpoC (Rv0668)   | 4063682 (Leu191Met:CTG→ATG) |
|          | LAM | Signal transduction mechanisms | 3239202 (Tyr654Asp: GTC→GTC) |
|          | X   | Carbohydrate transport and metabolism | 3352244 (Thr512Ala: ACC→CCC) |
|          | Ural| Intracellular trafficking, secretion, and vesicular transport | 3830349 (Ala391Thr: GGC→AGC) |
|          | AFR1| Intracellular trafficking, secretion, and vesicular transport | 2136642 (Leu191Met:CTG→ATG) |
|          | AFR2| Defense mechanisms | 1513189 (Ala48Val: GCA→GTA) |

Table 3. Genotype assignment of 34 MTB genomes using new proposed set of 10 SNPs.
Conclusion:
In conclusion, using the whole genome sequences of 249 MTB isolates, we identified a panel of 10 signature SNPs selected by combining IE criteria using spoligotyping and phylogenetic analysis and the selection of the nonsynonymous SNPs present in the most conserved cluster of orthologous groups (COG). This new proposed set of 10 SNPs can be used to build the molecular epidemiology of MTB, which can be incorporated in diagnostic assays and genotype-phenotype associations. The proposed system has the flexibility to be associated with the global databases where genotypes are based on the gold standard method: Spoligotyping. With the increase of genomes uploaded in global databases, future studies will obtain better support in the assignment of signature SNPs. The addition of the genomic signature along with drug resistance markers will improve TB control.

Acknowledgements:
This study was supported by the Peruvian National Institute of Health.

Author Contribution:
Conceived and designed the experiments: DT, HG. Performed the experiments: DT, LJ, VB. Analyzed the data: DT, LJ, VB, KL. Contributed materials/analysis tools: DT, LJ, HG. Wrote the paper: DT, LJ, VB, KL, MG, HG. Research as a Master’s thesis: LJ. All authors have approved the final manuscript. All authors read and approved the final manuscript.

Conflict of Interest:
The authors declare no conflicts of interest.

References:
[1] Cole ST et al. Nature. 1998 393(6685):537-44 [PMID: 9634230]
[2] Brudie K et al. BMC microbiology. 2006 6:23 [PMID: 16519816]
[3] Supply P et al. Mol Microbiol. 2000 36(3):762 [PMID: 10844663]
[4] Groenew PM et al. Mol Microbiol. 1993 10(5):1057 [PMID: 7934856]
[5] Kamerbeek J et al. J Clin Microbiol. 1997 35(4):907 [PMID: 9157152]
[6] van Embden JD et al. Journal of bacteriology. 2000 182(9):2393 [PMID: 10762237]
[7] Streicher EM et al. Journal of clinical microbiology. 2007 45(1):237 [PMID: 17065260]
[8] Van Soolingen D. J Intern Med. 2001 249(1):1 [PMID: 11168781]
[9] Mostrom P et al. Clin Microbiol Infect. 2002 8(11):694 [PMID: 12445006]
[10] Barnes PF & Cave MD. N Engl J Med. 2003 349(12):1149 [PMID: 13679530]
[11] Mokrousov I. Infect Genet Evol. 2008 8(6):777 [PMID: 18691674]
[12] Dasgupta K & Merzies D. The European respiratory journal 2005 25(6):1107 [PMID: 15929967]
[13] Iwamoto T et al. PLoS One. 2012 7(11):e94651 [PMID: 23185395]
[14] Goldblatt D et al. The international journal of tuberculosis and lung disease. 2014 18(9):1085 [PMID: 25189557]
[15] Odone A et al. European journal of public health. 2015 25(3):506 [PMID: 25500265]
[16] Zammarchi L et al. Infectious disease reports. 2014 6(4):5646 [PMID: 25568758]
[17] Dong H et al. Journal of clinical microbiology. 2010 48(11):4102 [PMID: 20739484]
[18] Desikan S & Narayanan S. Indian J Med Res. 2015 141(6):761 [PMID: 26205019]
[19] Brown AC et al. J Clin Microbiol. 2015 53(7):2230 [PMID: 25972414]
[20] Filliol I et al. Journal of bacteriology. 2006 188(2):759 [PMID: 16385065]

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 13(7): 224-230 (2017)
