Genomic determination of minimum multi-locus sequence typing schemas for *Mycoplasma hominis* to represent genomic phylogeny

Aleksey Jironkin[1], Rebecca J. Brown[1,2], Anthony Underwood[1], O. Brad Spiller[2], Victoria J. Chalker[1]
[1]Public Health England, London, UK; [2]Cardiff University School of Medicine, Cardiff, UK

Public Health England
Protecting and improving the nation's health

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

INTRODUCTION

Mycoplasma hominis is an opportunistic human pathogen and resides as a commensal on the mucosal surfaces of the cervix or vagina in 21 to 53% of sexually mature, asymptomatic women; and at a lower percentage in the urethra of males. Presence of *M. hominis* is associated with clinically diverse diseases including; urogenital diseases, postpartum fever, pneumonia, meningitis, post-operative wound infection, post-organ transplant infection, and septic arthritis. Although this organism has only been isolated from humans, the capacity of *M. hominis* to cause disease as a sole pathogen has been proven by induction of preterm labour and development of foetal chronic lung disease following experimental *in utero* administration of *M. hominis* to pregnant macaque monkeys.

Multi-locus sequence typing (MLST) analysis of the diversity of housekeeping genes that are considered to be under less selective pressure than other genes have been successfully employed for other mycoplasma species including *M. bovis [24]*, *M. agalactiae [25]*, *M. hyorhinis [26]* and *M. hyopneumoniae [27]*. Sogaard *et. al* examined six house-keeping gene sequences to investigate evidence of genomic recombination in *M. hominis* [28] and revealed a high degree of variability between these genes. However, the authors did not utilise the data to create a genotyping scheme.

The aim of this study is to develop a multi-locus sequence based typing scheme based on bioinformatics analysis to derive the minimum number of genes required to accurately reflect genomic phylogeny.

METHODS

- Isolates were assembled using Spades (v3.6.1) software with --only-assembler option and 21,33,45,53,65,77,83,93 kmers.
- Pan genome was constructed using NZ_CP009652.1, NC_013511.1, NZ_CP011538.1 and NZ_CP009677.1 references
- Pan genome is defined as a set of gene families with at least 1 allele found in one of the four references.
- Each gene family is constructed into a Hidden Markov Model using HMMER v3.2 software and combined into a database of models.
- Each isolate was scanned against the database of Hidden Markov Models to ascertain gene and allele composition.
- New alleles were added into the pan genome after scanning using the same method as clustering.
- For leave-one-out analysis, one gene was removed and a tree constructed using FastTree software. Topology was compared to the whole genome SNP topology derived by mapping each strain to ATCC27545 as reference and calling SNPs using GATK software.
- RAxML software was used to constructed Maximum-Likelihood phylogeny with GTR and GAMMA models.

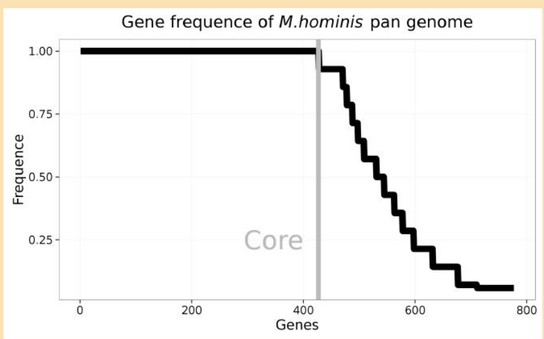

Figure 1. Pan-genome gene frequency across *M.hominis* isolates and reference genomes. 427 are conserved across all isolates

RESULTS

Raw reads from the samples were assembled and scanned against a database of Hidden Markov Models (HMMs) representing gene coding families constructed using four published reference genomes: NZ_CP009652 (533 genes), NC_013511 (497 genes), NZ_CP011538 (524 genes) and NZ_CP009677 (531 genes). On average, *M. hominis* pan genome clustering was able to detect mean 550 (median: 553) genes per sample, comparable to the mean number of genes found in the four reference genomes: 521. *M. hominis* pan genome contained total of 777 genes (Figure 1) with 427 genes (54.9%) present across all samples at least once and the ATCC27545 reference, a fraction similar to pan-genome size in other species. The shoulders in the pan-genome frequency distribution are likely to correspond to the genes found in the specific phylogenetic clades.

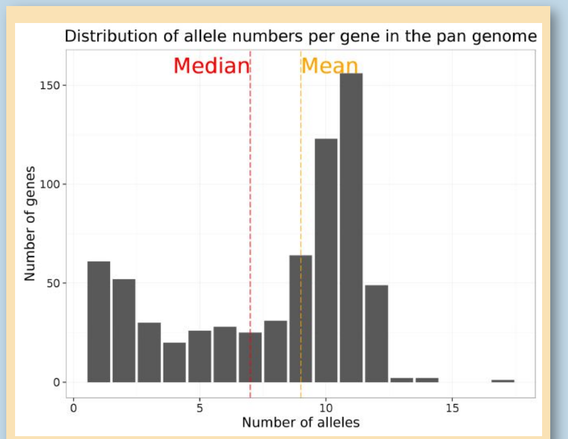

Figure 2. Distribution of the allele frequencies across all genes found in the pan genome of *M.hominis* shows 2 distinct peak associated with unique isolate and conserved genes.

A subset of genes found across all samples are likely to be highly conserved and carry little phylogenetic signal. Conversely, potentially small subset of genes could act as hotspots for the phylogenetic signal that drives tree topology. To find genes carrying the most signal, we have performed leave-one-out analysis: removing one gene at a time from the set of 427 core genes and constructing phylogenetic tree using remaining alleles from 426 genes. Resulting phylogeny is compared with the phylogeny derived using whole genome variants (gold standard). From the set of 427 genes, 379 genes conferred the same phylogenetic topology as the whole genome tree, remaining 48 genes disrupted the phylogenetic relationship of the samples to varying extent (Figure 2).

Although whole genome sequencing is growing in ubiquity with decreasing cost-per-sample, culturing *M. hominis* to produce enough biomass for whole genome sequencing is a challenge. As such, routine whole genome sequencing is not currently feasible, and typing using a technique that has a PCR amplification step, such as 7-locus MLST, is preferred. Although we have shown that the 48 genes are required to recapture the original phylogenetic relationships, we set out to find a set of genes that could be used in the traditional 7 locus MLST typing schema. To reduce the search space (73,629,072 total combinations) for a suitable combination of 7 genes, three sets of genes were selected: genes that caused lowest overlap with the whole genome tree in leave-one-out analysis (n=10), manually selected genes for their biological function (n=9) and combination of the above 2 sets (n=12). All 948 (120+36+792) possible combinations of the 3 sets of genes were analysed for the topology closest to the whole genome, 15 combinations have produced the highest topological similarity among all tested combinations. The resulting 15 trees were manually analysed for consistency and 3 schemas were found to confer the highest similarity to the whole genome tree. All 15 trees classified MH23 and MH28 into different subtree from the original whole genome analysis, with some trees also classifying MH17 differently. Selected 3 schemas had overall very similar topology to the original tree: 2 general clades and correct bifurcations. One schema correctly placed MH17 as outer most group to that clade, whereas others incorrectly placed MH17's ancestry. The branch length of the tree could not be reproduced as shorter sequences with different number of SNPs were used, compared to the original SNP sequence alignment. Overall the 3 schemas almost completely reproduced the phylogenetic relationship found using whole genome SNP data.

- Pan genome contains total of 777 genes
  - 427 genes were found to be core to the samples in this study
- Leave-one-out analysis resulted in the 379 genes producing invariant topology vs whole genome Maximum-Likelihood SNP tree constructed using RAxML
- 48 genes were found to alter the relationship between isolates
  - We propose this to be the minimum gene set for core genome MLST schema
- In order to find suitable 7 genes suitable for MLST schema we have reduced 73,629,072 possible combinations to 948 by manual selection
- 15 combinations produced equally close topology to the ML tree, albeit not identical
- 3 of 15 were chosen for highest topological similarity
- 3 7-locus MLST schemas consisted of total 9 genes.
- Average dN/dS ratio for 3 schemas was 0.11 and Hunter-Gaston average diversity index 0.965

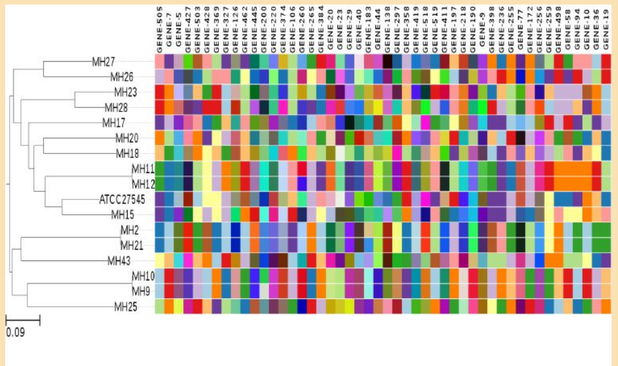

Figure 3. Minimal 48 gene core MLST set required to reproduce the phylogeny observed in the whole genome SNP tree. Colour in each column corresponds to an allele number.

DISCUSSION

Liu *et al.*[1] have looked at the *Mycoplasma* pan-genome and found only 196 genes to be conserved across the *Mycoplasma* specie, however the study did not include *M. hominis*. In contrast we have identified 427 genes to be part of the core genome, twice the previously reported number. To check potential congruence between the two core gene sets, we have used *Mycoplasma pneumoniae* M129(U00089) and *Ureaplasma parvum* serovar 3 (AF222894) genes to check for hits to our pan-genome set. There were no significant hits to any of the genes in our pan-genome set, suggesting that Hominis group is a distant relative to other *Mycoplasma* species and shares little sequence similarity in the coding genes.

To determine stability of the proposed genes for the MLST schemas 2 *M .hominis* strains have been passaged 10 times and sequenced to find any new gene variants. Genes found to have variation after 10 passages, would be unsuitable candidates for the MLST schema. From the set of 48 genes, one gene (*pcrA*) was found to have acquired mutations, and therefore is not suitable for use in a typing schema. This gene was not in the reduced 20 genes set and does not appear in the final MLST gene proposals (mosaic Figure 4). There could be potentially more unstable genes than we were able to detect using 2 strains.

Difficulty in obtaining biomass suitable for whole genome sequencing, means that there are few deposited whole genome sequences. 16 isolates used in this study, although adequate for inferring phylogenetic relationship, does not represent the full temporal and geographic diversity of *M. hominis* species. Due to low number of published isolates, it is difficult to validate the proposed schemas. However, this would improve as more isolates are sequenced and become available from online repositories.

Further work is required to experimentally validate the proposed schema for use within clinical setting

CONCLUSIONS

- In this study we have identified minimum gene set for core genome MLST schema consisting of 48 genes.
- The minimum gene set completely recapitulates the phylogenetic relationships seen in the whole genome SNP tree.
- After reducing total search space we have identified 3 sets of 7 genes most suitable to be part of a 7-locus MLST schema
- We have confirmed the stability of the proposed MLST gene sets with 10 passages from two strains

ACKNOWLEDGEMENTS

We would like to thank Public Health England and Cardiff University for funding the project, Public Health England for sequencing the isolates and various hospitals for providing isolates.

REFERENCES

1. W. Liu, L. Fang, M. Li, S Li, S. Guo, Z. Feng, B. Li, Z. Zhou, G. Shao, H. Chen and S. Xiao, "Comparative Genomics of Mycoplasma: Analysis of Conserved Essential Genes and Diversity of the Pan-Genome", *PLoS ONE*, p. e35698, 2012

Follow

@biocomputerist @PHE_uk