Novel SNP-based assay for genotyping *Mycobacterium avium* subsp. *paratuberculosis*

Celia Leão [a,b,#], Robert J Goldstone[c], Josephine Bryant[d] , Joyce McLuckie[a], João Inácio[e,f,#], David GE Smith[a,c] and Karen Stevenson[a]

Moredun Research Institute, Pentlands Science Park, Penicuik, United Kingdom[a]; Unidade Estratégica de Investigação e Serviços em Produção e Saúde Animal, Instituto Nacional de Investigação Agrária e Veterinária, I.P. (INIAV, I.P.), Lisboa, Portugal[b]; Institute of Infection, Inflammation & Immunity, University of Glasgow, Glasgow, United Kingdom[c]; Division of Infection and Immunity, University College London, London, United Kingdom[d]; Unidade de Microbiologia Médica, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal[e]; School of Pharmacy and Biomolecular Sciences, University of Brighton, Brighton, United Kingdom[f]

Running Head: **SNP assay for genotyping *M. a. paratuberculosis***

[#]Address correspondence to:
Célia Leão, celia.leao@iniav.pt
Instituto Nacional de Investigação Agrária e Veterinária, I.P. (INIAV, I.P.), Unidade Estratégica de Investigação e Serviços em Produção e Saúde Animal, Lisboa, Portugal
João Inácio, J.Inacio@brighton.ac.uk
School of Pharmacy and Biomolecular Sciences, University of Brighton, 815 Cockcroft building, Lewes road, Brighton BN2 4GJ, United Kingdom

ABSTRACT

Typing of *Mycobacterium avium* subspecies *paratuberculosis* strains presents a challenge since they are genetically monomorphic and traditional molecular techniques have limited discriminatory power. The recent advances and availability of whole genome sequencing has extended possibilities for the characterization of *Mycobacterium avium* subspecies *paratuberculosis* and it can provide a phylogenetic context to facilitate global epidemiology studies. In this study we developed a SNP assay based on polymerase chain reaction and restriction enzyme digestion or sequencing of the amplified product. The SNP analysis was performed using genome sequence data from 133 *Mycobacterium avium* subspecies *paratuberculosis* isolates with different genotypes from eight different host species and seventeen distinct geographic regions around the world. A total of 28402 SNPs were identified among all the isolates. The minimum number of SNPs required to distinguish between all the 133 genomes was 93 and between only the Type C isolates was 41. To reduce the number of SNPs and PCRs required we adopted an approach based on sequential detection of SNPs and a decision tree. By the analysis of 14 SNPs *Mycobacterium avium* subspecies *paratuberculosis* isolates can be characterized within 14 phylogenetic groups with a higher discriminatory power compared to MIRU-VNTR assay and other typing methods. Continuous updating of genome sequences are needed in order to better characterize new phylogenetic groups and SNP profiles. The novel SNP assay is a discriminative, simple, reproducible method and requires only basic laboratory equipment for the large-scale global typing of *Mycobacterium avium* subspecies *paratuberculosis* isolates.

INTRODUCTION

Mycobacterium avium subspecies *paratuberculosis* (*Map*) causes Johne's disease, a chronic infectious enteritis principally of ruminants. The disease occurs worldwide and is responsible for significant losses to the livestock industry. *Map* also has been detected in a subset of human patients with Crohn's disease (1) although the zoonotic role of the bacterium remains controversial.

Strain typing is a prerequisite for tracing the sources of infection and for studying the epidemiology, population structure and evolutionary relationships between isolates. It can also reveal the genetic diversity underlying important phenotypic characteristics such as host specificity, pathogenicity, antibiotic resistance and virulence. Typing of *Map* strains presents a challenge since *Map*, like *Mycobacterium tuberculosis*, is genetically monomorphic (2). Genetic diversity among *Map* strains has been investigated using molecular techniques such as restriction fragment length polymorphism and IS*900* analysis (IS*900* RFLP) (3), pulsed-field gel electrophoresis (PFGE) (4), amplified fragment length polymorphism (AFLP) analysis (5), random amplified polymorphic DNA (RAPD) analysis (6), mycobacterial interspersed repetitive unit – variable number tandem repeat (MIRU-VNTR) analysis (7) and short-sequence repeat (SSR) analysis (8). However, these techniques have limited discriminatory power when applied to *Map* and although this can be increased by combining complementary genotyping techniques, it is often insufficient for accurately determining relationships among isolates or global epidemiological studies (9; 10).

Whole genome sequencing (WGS) provides the ultimate resolution of isolates and, unlike the techniques above, it can provide a phylogenetic context to facilitate global epidemiology studies and affirm epidemiological connections (10; 11; 12). Although WGS is becoming cheaper, it is still too expensive to be used for routine genotyping of *Map* isolates and requires robust data handling and analysis processes. Single nucleotide polymorphisms (SNPs) have been used successfully to type several genetically monomorphic pathogens,

including *M. tuberculosis* (13), *Mycobacterium bovis* (14), *Salmonella enteritica* Typhi (15) and *Yersinia pestis* (16). SNP assays have been used to discriminate between *Map* strain Types I, II and III (17; 18) and between strains derived from animal and human hosts (19). However, these assays were based on a limited number of SNPs, many of which were not informative when applied to a wider wild-type population. The purpose of this study was to develop a SNP assay that is discriminative, practicable and reproducible for the large-scale global typing of *Map* isolates. Hence we developed a SNP typing method based on polymerase chain reaction (PCR) and restriction enzyme digestion, which would minimize costs and require only basic laboratory equipment. Additionally, we adopted an approach based on sequential detection of SNPs and a decision tree to reduce the number of SNPs and PCRs required.

MATERIALS AND METHODS

Selection of genome sequences and *Map* strains. Genome sequences from 133 *Map* isolates generated in a previous study (10) were selected for SNP and phylogenetic analyses. This panel was chosen to maximise genetic diversity and reduce phylogenetic discovery bias (2) since it comprised isolates with different genotypes (determined by multiplex PFGE and MIRU-VNTR), which were selected from 17 different countries and isolated from eight different host species. The panel also included isolates representing the major strain types that have been described previously (reviewed by 20; 21). The sequence reference numbers are given in Figures 1 and 2. The field isolates (n=26) for which genome sequence data was not available and that were used to validate the SNP assay are shown in Table 1.

Preparation of genomic DNA. Genomic DNA from the UK field isolates (n=10) (Table 1) was extracted from plugs used to perform the PFGE analysis. Briefly, half of each plug was washed three times with 2 ml of Tris EDTA buffer pH 8 (10 mM Tris.Cl, 1 mM EDTA) for

10 min with shaking. After washing the plugs, 0.5-1 ml of sterile deionised water was added and the agarose was melted at 70ºC. The suspension was stored at -20ºC until required for PCR. DNA from the Portuguese field isolates (n=16) (Table 1) was extracted from pure cultures grown in Middlebrook 7H9 medium supplemented with 10% OADC (Oleic Albumin Dextrose Catalase) and 2 mg/l of mycobactin as previously described (10) using the QIAamp DNA Mini Kit (Qiagen) according to manufacturer's instructions with minor modifications. Briefly, the bacterial suspension was centrifuged at 5000 × g for 10 min and the pellet was resuspended in 180 µl of ATL buffer. Zirconium beads (0.1 mm) were added to the mixture and mechanical disruption of the cells was performed twice with a FastPrep FP120 Bio101 bead shaker (Savant Instruments Inc., Holbrook, NY) at 6.5 $msec^{-1}$ for 45 seconds. The disrupted mixture was cooled on ice for 15 min and 20 µl of proteinase K was added. The remaining procedure was completed according to manufacturer's instructions. The DNA was eluted with 200 µl of AE buffer and stored at -20ºC until required for PCR.

SNP analysis and phylogenomics. The genome sequence of *Map* K10 (22; Accession number NC_002944.2) was used as the reference genome for the phylogenetic analyses. Raw genome sequence data for the *Map* isolates are available from the European Nucleotide Archive under accession PRJEB2204. The assembly of the reads into contigs was performed using Velvet assembler program (freely available at https://www.ebi.ac.uk/~zerbino/velvet/) and the alignment of the sequences and positioning of SNPs was executed utilising MUMmer package (freely available at http://mummer.sourceforge.net/) using *Map* K10 as reference sequence. SNPs were then extracted and concatenated and a phylogenetic tree was calculated using phyML (freely available at http://atgc.lirmm.fr/phyml). This phylogenetic tree was then imported into R using the ade4 package to explore the 'paths' which exist between the genomes. These 'paths' are a description of the strains which group on the same branch as the structure of the phylogenetic tree descends (i.e. all strains are included on the 'root branch',

which then bifurcates to include a subset of strains on one branch and the remaining strains on the other, and so on). Using these 'paths' to assign groups of genomes, SNPs which were shared by all strains on each branch were compiled. This data then allowed comparison of the SNPs present in reducing sized groups of genome sequences (as the tree descends through specific branches, the number of genomes remaining in each subsequent group becomes progressively smaller). This data permitted the detection of discriminating SNPs which are present in one group of sequences yet absent in others. A 'set cover' analysis was then performed on this dataset to calculate the minimum number of SNPs which were required to discriminate between the groups of strains at each selected level of the tree (see below). These SNPs were then selected and taken forward for validation as below. SNPs were named according to their base position in the revised *Map* K10 genome sequence.

Set cover approximation and decision tree construction. The set cover problem asks, given a universe of elements to be covered (in this case, the universe was a series of pair-wise combinations of all genome groups), and a number of 'sets' which contain varying combinations of the elements to be covered (in this case, these 'sets' are specific SNPs which are found to discriminate between one or more of the pair-wise combinations of genome groups), what is the minimum number of sets required to cover all the elements of the universe. In this case, the set cover asks what is the minimum number of SNPs required to 'fill' all the pairwise combinations of genome groups. In this way, the defined set of SNPs would discriminate between every possible pairwise combination of genome groups. The set cover problem is a class of decision problem known to be NP-complete, meaning the time taken to reach an exact solution increases exponentially with problem size. Due to the large number of total SNPs and genome groups to consider, we opted to use an approximation algorithm known as the greedy solution to estimate the set cover. Greedy algorithms work by iteratively selecting the set which covers the largest number of uncovered universe elements.

The greedy approach is considered a suitable polynomial time approximation for the set cover problem. This analysis resulted in the identification of a set of SNPs which could distinguish between all pairwise combinations of genome groups.

The chosen SNPs which comprised the set cover formed the basis of the decision tree for determining group assignment of genomes, since every bifurcation of the phylogenetic tree could be discriminated by at least one SNP present in the set cover. The structure of the phylogenetic tree was then manually investigated in light of the SNPs within the set cover to minimise the number of SNPs required to determine between phylogenetically relevant branches. These SNPs were then validated as outlined below.

SNP validation and primer design. Primers for PCR amplification of the selected SNPs were designed using the online software Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/) and the revised *Map* K10 genome sequence (23; Accession number AE016958.1) (Table 2). Primers were designed to be 18-20 mer with a melting temperature between 63 and 67ºC.

PCR reactions were carried out in 50 µl containing 200 µM of each deoxynucleotide triphosphate (Invitrogen), 0.5 µM of each primer (Table 2), 1 U of Phusion® High- Fidelity DNA polymerase, 1× Phusion GC buffer (New England Biolabs), and 4 µl of extracted DNA (5 ng/µl). Amplification was performed in a TC-PLUS Thermal cycler (Techne) with an initial step at 98 ºC for 3 min, followed by 35 cycles at 98 ºC for 30 sec, 63-67 ºC for 30 sec (annealing temperatures provided in Table 2) and 72 ºC for 40 sec, ending with a step at 72 ºC for 10 min. The amplified products were electrophoretically analysed in a 1.5% (w/v) agarose gel stained with SYBR® Safe DNA Gel Stain (Life Technologies) in 0.5× Tris-Borate-EDTA (TBE) buffer. Gel electrophoresis images were acquired with an Alphaimager™ 2200 (Alpha Innotech). DNA ladder IV (Bioline) and *Map* K10 (positive control) were included on each gel.

Restriction endonuclease analysis of PCR products was performed according to the manufacturer's instructions using one unit of restriction enzyme and 10 µl of amplified product in a total reaction volume of 25 µl. All restriction endonucleases were purchased from New England Biolabs (Table 2). Restricted products were detected by electrophoresis on 1.5% (w/v) agarose gels as described above.

Confirmation of the presence of the SNPs was obtained by sequencing the PCR products. PCR product (40µl) was purified using QIAquick PCR purification Kit (Qiagen) according to manufacturer's instructions. Sequencing of PCR products was carried out by Eurofins Genomics (MWG-Biotech) using the same primers used to the amplification of the fragments. Confirmation of the presence or absence of the SNP in the expected position of the genome was achieved using Basic Local Alignment Search Tool (Blast - http://blast.ncbi.nlm.nih.gov/Blast.cgi). The phylogenetic profile for each isolate was obtained by combining the results for all SNPs.

Discriminatory power of SNP-based genotyping assay. The discriminatory power of the assay was calculated using the Hunter-Gaston Discriminatory Index (HGDI), according to Hunter and Gaston (24).

RESULTS

The genome sequence data from 133 *Map* strains were compared to the reference *Map* strain K10 to identify SNPs. A total of 28402 SNPs were identified among all the isolates. A phylogenetic tree was generated based on the SNP analysis and distinct phylogenetic groups were identified (Figures 1 and 2), which conformed to the broadly recognised phylogenetic structure of *Map* (10). By using an adaptation of the 'set cover' problem, the minimum number of SNPs required to discriminate between all the isolates was calculated to be 93. To

refine the number of SNPs to a number manageable for routine laboratory procedures, we considered a strategy based on sequential detection of SNPs and a decision tree (Figure 3). The phylogenetic analysis distinguished two major strain groups corresponding to those previously designated Type C and Type S (Figure 1). We identified a SNP (snp3842359), which could be detected using BsmB1 (Table 2) that would discriminate between these two groups. Type C strains have an 'A' and Type S strains either a 'G' or 'C' at base position 3842359. This constituted the first step in the decision tree, the next step being determined according to whether the isolate was Type C or Type S (Figure 3).

For further analysis of Type S strains, we identified a SNP (snp343677), which could be detected using AvaII that discriminated the Type S sub-groups Type I and Type III (Table 2, Figures 1 and 3). Additionally, snp3842359 also could be used to distinguish Type I and Type III strains since Type I have a "G" and Type III have a "C" at base position 3842359 (Figure 3), which could be detected by PCR amplification and sequencing of the product.

The Type C group comprised the majority of the isolates and the high homogeneity within this group posed a challenge for identification of clade specifying SNPs. Firstly, SNP analysis was repeated with only the sequence data from the 115 Type C isolates (Figure 2) and the minimum number of SNPs required to distinguish between these 115 isolates was determined to be 41. We then considered three principal sub-groups designated Bison (as reported previously, 10) and A and B as shown in Figures 1 and 2. We identified a SNP (snp50173), which could distinguish the Bison group from both sub-groups A and B using ApoI (Table 2, Figures 1 and 3). This constituted step 2 in the decision tree (Figure 3). Within the Bison group, Indian Bison type could be differentiated from US Bison type using snp2327379 and further differentiation of the Indian bison type was possible using snp305277 (Table 2, Figure 3). Due to the limited number of Bison type strains available extensive verification of these SNPs was not possible in this study.

To further discriminate between Type C isolates in sub-groups A and B, we identified a SNP (snp4111202), which could be detected using FatI (Table 2, Figures 1 and 3). Due to the small number of isolates in sub-group B, we did not, at this stage, seek additional SNPs to further discriminate the isolates within this group. This constituted step 3 in the decision tree (Figure 3).

Within the larger subgroup A, SNPs were identified that could discriminate ten groups (snp3879247, snp2939977, snp1932058, snp1327872, snp3844632, snp1966028, snp305277, snp4339946, snp2087274 and snp1686154, Tables 2 and 3, Figures 2 and 3). These SNPs were verified by sequencing of the PCR products and comparison of the sequences with the reference *Map* K10 strain. It was not possible to identify SNPs with specific restriction endonuclease sites for the clades differentiated using snp2939977 and snp1932058 but all other SNPs could be detected by restriction endonuclease analysis (Table 2).

SNP validation and genotyping. For the validation of the selected SNPs comprising the decision tree (Figure 3), DNA from isolates belonging to Type C (Bison group: MAPMRI029, MAPMRI031, MAPMRI127, MAPMRI034, MAPMRI117 and MAPMRI026; Sub-group A: MAPMRI110, MAPMRI120 and MAPMRI027; Sub-group B: MAPMRI059, MAPMRI136 and MAPMRI091) and Type S (Type I: MAPMRI007and MAPMRI001; Type III: MAPMRI051, MAPMRI045 and MAPMRI047) (Figure 2) were used for amplifying products containing SNPs (as indicated in Figure 3) and the PCR products were digested with the correspondent restriction enzymes (Table 2). PCR products were also purified and sequenced to confirm SNPs. To assess the validity of the ten SNPs for discriminating the clades within Sub-group A, thirty isolates from the original panel of 115 sequenced strains were re-tested. These were selected to be representative of the ten different phylogenetic clades as shown in Figure 2 and were subjected to analyses for all 14 SNPs (Figure 3), all of

252 which were confirmed to be present. The *Map* isolates were grouped into SNP profiles as
253 shown in Table 3 and Figure 3.
254 A further 26 *Map* isolates (Table 1) not previously sequenced or typed using this SNP assay
255 were genotyped using the 14 SNPs. These isolates belonged to four phylogenetic groups
256 within the Sub-group A. Significantly, ten isolates from different geographic regions of
257 United Kingdom with the same MIRU-VNTR and PFGE profile were classified into three
258 different SNP profiles, one of which was not identified in the original phylogenetic analysis
259 and therefore represented a new SNP profile (SNP11) (Table 3). Two UK isolates were
260 identified to belong to the SNP1 profile, six isolates to SNP9 and two isolates to profile
261 SNP11. The 16 Portuguese isolates were distributed among three phylogenetic groups: all the
262 isolates from Azores were present in the same group identified by profile SNP3; two isolates
263 from the same region in the north of Portugal were identified in the same phylogenetic group
264 as six isolates from the United Kingdom (profile SNP2); and, the remaining four Portuguese
265 isolates from the same region were found to belong to the new phylogenetic group together
266 with the two DNAs from the United Kingdom (profile SNP11) (Table 1).
267
268 **Discriminatory power of SNP-based genotyping assay.** The discriminatory power of the
269 SNP-based assay was 0.8390 for the 56 isolates that were used for the validation of the assay.
270 In order to compare the discriminatory power of the SNP assay with MIRU-VNTR analysis,
271 we used the typing results for 46 isolates, which had been typed using both methods. The
272 HGDI was calculated to be 0.8135 for the SNP assay and 0.6386 for MIRU VNTR.
273
274 **DISCUSSION**
275 Several methods have been used to characterize *Map* strains but they have some limitations.
276 Techniques based on the analysis of total genomic DNA such as RFLP and PFGE require
277 culture of the isolates to prepare moderate amounts of high quality DNA and are therefore

slow, can be technically demanding, labour intensive, hard to standardize and expensive. Furthermore, RFLP and PFGE can clearly distinguish between Type C and S but do not give sufficient discrimination within these strain types for detailed epidemiological studies. Techniques such as AFLP and RAPD employ PCR to detect smaller genomic DNA fragments but are less utilised for epidemiological studies due to difficulties in standardisation, reproducibility and limited discriminative power. Other typing methods based on repetitive sequences such as SSR and MIRU-VNTR are popular due to their ease of use and rapidity but are again limited with respect to their ability to discriminate within the two major strain types and the typing results may not reflect the evolutionary relationships between isolates (10; 12; 25; 26).

In this study a novel typing assay based on SNP analysis by PCR and restriction or sequencing of the amplified products was developed. This technique is easy to perform, is applicable to a small quantity of genomic DNA and is based on standard PCR and restriction endonuclease analysis. It was possible to refine the number of SNPs to a number manageable for routine laboratory procedures by adopting an approach based on sequential detection of SNPs via a decision tree. The SNP assay was highly discriminative, possessing a higher discriminatory power than MIRU-VNTR when applied to 46 *Map* isolates.

SNP-based typing assays are particularly useful for monomorphic pathogens that exhibit limited genetic diversity. Furthermore, they have the advantage that they can be used to determine phylogenetic relationships, unlike techniques based on mobile or repetitive DNA elements, which interrogate a relatively small proportion of the mycobacterial genome and can exhibit homoplasy. However, SNP discovery is subject to phylogenetic discovery bias (2), a phenomenon well described for *M. tuberculosis* (27) and *Bacillus anthracis* (28), and is most likely to be encountered where information is missing on strains geographically-restricted or belonging to rare phylogenetic groupings. For this reason, we utilised a large collection of global isolates, which had been previously genotyped by classical molecular

304 tools (PFGE and MIRU-VNTR) to maximise genetic diversity and include representatives of
305 all previously reported strain types. The SNPs identified in this study should provide the
306 necessary means to unambiguously classify *Map* strains within this global framework. Even
307 so, the composition of any panel of SNPs needs to be reviewed or augmented once additional
308 groups of strains that were not included in the initial analysis are discovered. This has been
309 illustrated in this study with the discovery of a new phylogenetic group represented by profile
310 SNP 11 comprising six isolates when an additional uncharacterised 26 isolates were screened
311 using the SNP assay. WGS needs to be performed and the sequence comparisons and SNP
312 analysis repeated to determine the positions of these isolates within the phylogenetic tree and
313 determine any additional SNPs that could be used to define the group. In this study, we
314 concentrated on finding SNPs to differentiate within Type C strains in sub-group A. The SNP
315 assay needs to be expanded to differentiate between strains within Type S, Bison type and
316 Type C subgroup B, but WGS data from more strains belonging to these phylogenetic groups
317 are required for SNP discovery to provide a phylogenetically robust framework for strain
318 differentiation combined with sufficient discriminatory power for detailed genetic studies.
319 SNPs have been described in previous studies that differentiate between the two major
320 phylogenetic groups, Type S and Type C, and between *Map* strain Types I and III. A PCR-
321 REA assay described by Whittington et al. (29) based on a SNP at base position 223 in the
322 IS*1311* insertion sequence has been used extensively for discriminating between Type S and
323 Type C strains. However, in a recent study (10) the IS*1311*-REA incorrectly identified strain
324 MAPMRI074 as a Type S strain when WGS and SNP analysis clearly confirmed it to be
325 Type C, suggesting that the C to T allelic variation at base pair position 223 in IS*1311*
326 occurred after the initial divergence of Type C from Type S strains. The SNP identified in
327 this study, snp3842359, and the corresponding restriction endonuclease BsmBI for PCR-REA
328 SNP detection could provide a more reliable alternative assay for differentiating Type S and
329 C strains.

IS*1311* SNP analysis has also been used to distinguish Bison type strains from non-Bison Type C and Type S strains (30). In Bison-type strains, all copies of IS*1311* have a "T" at base pair position 223, whereas the non-Bison Type C strains have one or more copies with a "C" or "T" at the same position. Copy number with respect to this allele is not always easy to assess and can be very variable (10). Snp50173 and the corresponding restriction endonuclease ApoI for PCR-REA SNP detection could provide an easier, alternative assay for discriminating Bison-type strains from other Type C strains.

A study published by Castellanos and colleagues (17) developed a PCR-REA to detect a SNP present on the *gyrB* gene at base position 1626 that allowed discrimination of Type III from Type I and II strains. Additional SNPs in the *gyrA* gene were identified that could differentiate Types I and III from Type II. In our study it is possible to use only a single SNP to discriminate Type I, Type II and Type III based on the amplification and sequencing of a fragment containing the snp3842359 where in the same position of the genome Type I strains have a "G", Type II have an "A" and Type III have a "C". This is an improvement compared with the system previously reported.

In conclusion, we developed a novel SNP-based genotyping assay based on the analysis of 14 SNPs that can be used to characterize *Map* isolates within 14 phylogenetic groups with a higher discriminatory power compared to MIRU-VNTR assay and other typing methods. We adopted an approach based on sequential detection of SNPs and a decision tree based on PCR-restriction enzyme digestion to reduce the number of SNPs and required PCRs. This novel assay can overcome some issues regarding the genotyping of isolates characterized as Type I, Type III and Bison type. Continuous updating of genome sequences are needed in order to better characterize new phylogenetic groups and SNP profiles. The novel SNP assay is a discriminative, simple, reproducible method and requires only basic laboratory equipment for the large-scale global typing of *Map* isolates.

Acknowledgements

This work was supported by the Scottish Government Rural and Environment Science and Analytical Services Division. Célia Leão is a recipient of a PhD grant from "Fundação para a Ciência e a Tecnologia" SFRH/BD/62469/2009.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

References

1. **Naser SA, Sagramsingh SR, Naser AS, Thanigachalam S**. 2014. *Mycobacterium avium* subspecies *paratuberculosis* causes Crohn's disease in some inflammatory bowel disease patients. World J Gastroenterol **20**:7403-7415.
2. **Achtman M.** 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Ann Rev Microbiol **62**:53-70.
3. **Pavlik I, Horvathova A, Dvorska L, Bartl J, Svastova P, du Maine R, Rychlik I.** 1999. Standardisation of restriction fragment length polymorphism analysis for *Mycobacterium avium* subspecies *paratuberculosis*. J Microbiol Meth **38**:155-167.
4. **Stevenson K, Hughes VM, de Juan L, Inglis NF, Wright F, Sharp JM**. 2002. Molecular characterization of pigmented and non-pigmented isolates of *Mycobacterium avium* subspecies *paratuberculosis*. J Clin Microbiol **40**:1798-1804.
5. **Motiwala AS, Strother M, Amonsin A, Byrum B, Naser SA, Stabel JR, Shulaw WP, Bannantine JP, Kapur V, Sreevatsan S.** 2003. Molecular epidemiology of *Mycobacterium avium* subsp. *paratuberculosis*: Evidence for limited strain diversity, strain sharing, and identification of unique targets for diagnosis. J Clin Microbiol **41**:2015-2016.

6. **Pillai SR, Jayarao BM, Gummo JD, Hue Jr EC, Tiwari D, Stabel JR, Whitlock RH.** 2001. Identification and sub-typing of *Mycobacterium avium* subsp. *avium* by randomly amplified polymorphic DNA. Vet Microbiol **79**:275-284.
7. **Thibault VC, Grayon M, Boschiroli ML, Hubbans C, Overduin P, Stevenson K, Gutierrez MC, Supply P, Biet F.** 2007. New variable-number tandem-repeats markers for typing *Mycobacterium avium* subsp. *paratuberculosis* and *M. avium* strains: comparison with IS*900* and IS*1245* restriction fragment length polymorphism typing. J Clin Microbiol **45**:2404-2410.
8. **Sevilla I, Li L, Amonsin A, Garrido J, Geijo MV, Kapur V, Juste RA**. 2008. Comparative analysis of *Mycobacterium avium* subsp. *paratuberculosis* isolates from cattle, sheep and goats by short sequence repeat and pulsed-field gel electrophoresis typing. BMC Microbiol **8**:204.
9. **Stevenson K, Àlvarez J, Bakker D, Biet F, de Juan L, Denham S, Dimareli Z, Dohmann K, Gerlach G-F, Heron I, Kopecna M, May L, Pavlik I, Sharp JM, Thibault VC, Willemsen P, Zadoks R, Greig A.** 2009. Occurrence of *Mycobacterium avium* subspecies *paratuberculosis* across host species and European countries with evidence for transmission between wildlife and domestic ruminants. BMC Microbiol **9**:212.
10. **Bryant JM, Thibault VC, Smith DGE, McLuckie J, Heron I, Sevilla IA, Biet F, Harris SR, Maskell DJ, Bentley SD, Parkhill J, Stevenson K**. 2015. Phylogenomic exploration of the relationships between strains of *Mycobacterium avium* subspecies *paratuberculosis.* BMC Genomics In press
11. **Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SA, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van Soolingen D.** 2013. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis **13**:100.

12. **Ahlstrom C, Barkema HW, Stevenson K, Zadoks RN, Biek R, Kao R, Trewby H, Haupstein D, Kelton DF, Fecteau G, Labrecque O, Keefe GP, McKenna SLB, Buck JD**. 2015. Limitations of variable number of tandem repeat typing identified through whole genome sequencing of *Mycobacterium avium* subsp. *paratuberculosis* on a national and herd level. BMC Genomics **16**:161.
13. **Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, Borrell S, Fenner L, Comas I, Coscollà M, Gagneux S.** 2012. Two new rapid SNP-Typing methods for classifying *Mycobacterium tuberculosis* Complex into the main phylogenetic lineages. PLOS ONE **7**:e41253.
14. **Joshi D, Harris NB, Waters R, Thacker T, Mathema B, Krieswirth B, Sreevatsan S.** 2012. Single Nucleotide Polymorphisms in the *Mycobacterium bovis* genome resolve phylogenetic relationships. J Clin Microbiol **50**:3853-3861.
15. **Octavia S, Lan R.** 2007. Single-Nucleotide-Polymorphism typing and genetic relationships of *Salmonella enterica* Serovar Typhi Isolates. J Clin Microbiol **45**:3795–3801.
16. **Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M.** 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet **42**:1140-1143.
17. **Castellanos E, Aranaz A, Romero B, de Juan L, Àlvarez J, Bezos J, Rodríguez S, Stevenson K, Mateos A, Domínguez L.** 2007. Polymorphisms in *gyrA* and *gyrB* genes among *Mycobacterium avium* subsp. *paratuberculosis* Type I, II, and III isolates. J Clin Microbiol **45**:3439-3442.

18. **Gastaldelli M, Stefani E, Lettini AA, Pozzato N.** 2011. Multiplexed typing of *Mycobacterium avium* subsp. *paratuberculosis* Type I, II and III by Luminex xMAP suspension array. J Clin Microbiol **49**:389-391.

19. **Wynne JW, Bellera C, Boyda V, Francisc B, Gwoździd J, Carajiasd M, Heinea HG, Wagnere J, Kirkwoode CD, Michalskia WP.** 2014. SNP genotyping of animal and human derived isolates of *Mycobacterium avium* subsp. *paratuberculosis*. Vet Microbiol **172**:479-485.

20. **Stevenson K. 2010**. Comparative Differences between Strains of *Mycobacterium avium* subsp. *paratuberculosis* p126-137. *In* Behr MA, Collins DM (eds), Paratuberculosis Organism, Disease, Control. CAB International, Oxfordshire, UK.

21. **Stevenson K. 2015**. Genetic diversity of *Mycobacterium avium* subsp. *paratuberculosis* and the influence of strain type on infection and pathogenesis: A review. Vet Res **46**:64.

22. **Li L, Bannantine JP, Zhang Q, Amonsin A, May BJ, Alt D, Banerji N, Kanjilal S, Kapur V**. 2005. The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. Proc Natl Acad Sci USA **102:**12344-12349.

23. **Wynne JW, Seeman T, Bulach D, Coutts SA, Talaat AM, Michalski WP**. 2010. Re-sequencing the *Mycobacterium avium* subspecies *paratuberculosis* K10 genome: improved annotation and revised genome sequence. J Bacteriol **192**:6319-6320.

24. **Hunter PR, Gaston MA.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol **26**:2465–2466.

25. **Castellanos E, de Juan L, Domínguez L, Aranaz A**. 2012. Progress in molecular typing of *Mycobacterium avium* subspecies *paratuberculosis*. Res Vet Sci **92**:169-179.

26. **Sevilla I, Garrido JM, Geijo M, Juste RA.** 2007. Pulsed-field gel electrophoresis profile homogeneity of *Mycobacterium avium* subsp. *paratuberculosis* isolates from cattle and heterogeneity of those from sheep and goats. BMC Microbiol 7:18.

27. **Gagneux S, Small PM.** 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis **7**:328-337.

28. **Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN, Huynh LY, Fraser CM, Keim P.** 2004. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci USA **101**:13536-13541.

29. **Whittington RJ, Marsh IB, Choy E, Cousins D.** 1998. Polymorphisms in IS*1311*, an insertion sequence common to *Mycobacterium avium* and *M. avium* subsp. *paratuberculosis*, can be used to distinguish between and within these species. Mol Cell Probe **12**:349-358.

30. **Whittington RJ, Marsh IB, Whitlock RH.** 2001. Typing of IS*1311* polymorphisms confirms that bison (*Bison bison*) with paratuberculosis in Montana are infected with a strain of *Mycobacterium avium* subsp *paratuberculosis* distinct from that occurring in cattle and other domesticated livestock. Mol Cell Probes **15**:139-145.

Figure 1. Whole genome SNP-based phylogenetic tree of 133 *Map* isolates included in this study (strain sequence reference MAPMRI numbers are indicated). Previously described lineages and sub-groups A and B described in this study are highlighted in grey.

Figure 2. Whole genome SNP-based phylogenetic tree of 115 Type C *Map* isolates (strain sequence reference MAPMRI numbers are indicated). Ten clades within the phylogenetic sub-group A can be distinguished by PCR and sequencing following the analysis of ten SNPs (grey boxes). Black circles indicate the 30 strains used for the validation of the method.

Figure 3. Work flow with a schematic representation of the decision tree with the sequential numbered steps and restriction enzymes, SNPs positions, expected bases and SNP profiles obtained based on the SNPs analysis. * SNP that can be detected only by sequencing of the amplified product. # Bison type strains from India. § Bison type strains from US.

Table 1. Additional *Map* field isolates used for validation of SNP assay

Isolate	Host	Geographic location	Multiplex PFGE profile	INMV profile*	SNP profile (this study)
F043531	cattle	Northern Ireland (UK)	2.1	1	1
F012398	cattle	Cumbria (UK)	2.1	1	11
F005713	cattle	Wiltshire (UK)	2.1	1	11
C217551	cattle	Selkirkshire (UK)	2.1.	1	9
C216785/2	cattle	Mid Glamorgan (UK)	2.1	1	9
C219376	cattle	Shropshire (UK)	2.1	1	9
C221325	cattle	Gwynedd (UK)	2.1	1	9
C524656	cattle	Aberdeenshire (UK)	2.1	1	1
C326442	cattle	Ayr (UK)	2.1	1	9
C216962/6	cattle	Leicestershire (UK)	2.1	1	9
C1	goat	São Miguel (Azores, PT)	ND	ND	3
C2	goat	São Miguel (Azores, PT)	ND	ND	3
C4	goat	São Miguel (Azores, PT)	ND	ND	3

C7	goat	São Miguel (Azores, PT)	ND	ND	3
C9	goat	São Miguel (Azores, PT)	ND	ND	3
C11	goat	São Miguel (Azores, PT)	ND	ND	3
C13	goat	São Miguel (Azores, PT)	ND	ND	3
C14	goat	São Miguel (Azores, PT)	ND	ND	3
C16	goat	São Miguel (PT)	ND	ND	3
C4A4	goat	São Miguel (Azores, PT)	ND	ND	3
B1	cattle	Vila do Conde (PT)	ND	2	9
B3	cattle	Vila do Conde (PT)	ND	2	9
B13	cattle	Póvoa do Varzim (PT)	ND	2	11
B18	cattle	Póvoa do Varzim (PT)	ND	2	11
B21	cattle	Póvoa do Varzim (PT)	ND	2	11
B22	cattle	Póvoa do Varzim (PT)	ND	2	11

ND - not determined; UK – United Kingdom; PT – Portugal. *Profile number assigned by INRA and published in an online database http://mac-inmv.tours.inra.fr

Table 2. PCR primers and restriction endonucleases used in the SNP assay for this study

SNP name	Primer name	Primer sequence (5' to 3')	Annealing temperature (^{0}C)	Product size (bp)	Enzyme	Restriction results§
snp3842359*	MAP_F1	CACCTGGCCAAGTACTACCA	63	528	BsmBI	(A) Type C - 528 bp
	MAP_R1	GCGATGTCATGATGCTGCTG				(G/C) Type S - 312, 216 bp
snp343677	TypeS_F	AACACCAGGATCGCGTTCTT	65	511	AvaII	(G) Type I - 297, 152, 62 bp
	TypeS_R	CAATTAGCGGTCGAGTCGTC				(A) Type III - 293, 218 bp
snp50173	BisonF	GGACGATTACTCGGTTCCAG	63	469	ApoI	(T) Bison group - 226, 192, 51 bp
	BisonR	ACCCGTGTTCGGCTACCT				(G) Cattle group - 277, 192 bp
snp4111202	SNP4_F	GTCAGAAACATCCCGCCTTC	65	461	FatI	(G) Type C sub-group A – 284, 177 bp
	SNP4_R	GTATTGAGTGAGGCAAGCGG				(C) Type C sub-group B - 461 bp
snp3879247	SNP5_F	GTTGATCGACAGCGAGTGC	64	465	BlpI/DdeII	(C) – 227, 238 bp
	SNP5_R	GTGGTGTCCGAGGTGAACTT				(T) – 465 bp
snp2939977	SNP6_F	TATCTCCAAGGACGCATTCC	64	516	-	-
	SNP6_R	CTGCCATGTCCGTCCTTAAT				
snp1932058	SNP7_F	GGCTTGAAACTCCAACTGCT	63	452	-	-
	SNP7_R	CGTCGTACATCCTCGTGGT				
snp1327872	SNP8_F	GCGCTTGTTGTACAGGTTGA	64	528	AvaI/BsoBI	(G) – 292, 171, 65 bp
	SNP8_R	TACGACGAAGACCCCGACTA				(T) – 463, 65 bp
snp3844632	SNP9_F	GATCGATGCGGAGCTCGT	64	457	FatI	(G) – 457 bp
	SNP9_R	TGACAGGAAGGTCCATAGCC				(C) – 241, 217 bp
snp1966028	SNP10_F	GTCGAGGGCTTCCAGGTT	67	427	SapI/EarlI	(A) – 427 bp
	SNP10_R	GTCTGAGGCCAGCGACAC				(C) – 246, 182 bp
snp305277†	SNP11_F	CCATCCCGAGTTCAACAAGT	64	461	BspMI/BfuAI	(G) – 310, 151 bp
	SNP11_R	ACTTGTCGGGGTTGTAGCTG				(A) – 461 bp
snp4339946	SNP12_F	AACCGCTCAAGGCGAAAG	64	464	BstEII	(T) – 292, 172 bp
	SNP12_R	TCCCTTATCTGCGAAGTGCT				(A) – 464 bp
snp2087274	SNP13_F	CAGACCGAGCACCTCCTG	65	453	HpyAV	(C) – 453 bp
	SNP13_R	CCGCGTTGAAGGATCTCAAG				(A) – 227, 226 bp
snp1686154	SNP14_F	GAATCCCCGGAACTGGTG	65	525	MscI	(G) – 525 bp

SNP14_R	GCAGTCCAGATAACGGAACG	(A) – 284, 241 bp

*SNP can also distinguish between Type I and III (Type I has a ”G” and Type III has a “C” at base position 3842359 detectable by sequencing the PCR product); †SNP can also distinguish between two phylogenetic subgroups of Bison group isolates (Figure 3); § in parenthesis is the expected base at the correspondent SNP position; - not applicable.

Table 3. SNP profiles of Type C *Map* isolates in phylogenetic sub-group A used in this study

Phylogenetic group	SNP profile	No. isolates verified	Base at SNP position*													
			3842359	343677	50173	4111202	3879247	2939977	1932058	1327872	3844632	1966028	305277	4339946	2087274	1686154
Reference base (K10)			A	A	G	G	C	G	G	G	G	A	G	T	C	A
Clade 1	1	7	A	A	G	G	C	G	**A**	G	G	A	G	T	C	G
Clade 2	2	1	A	A	G	G	C	**A**	G	G	G	A	G	T	C	G
Clade 3	3	11	A	A	G	G	**T**	G	G	G	G	A	G	T	C	G
Clade 4	4	2	A	A	G	G	C	G	G	**T**	G	A	G	T	C	G
Clade 5	5	1	A	A	G	G	C	G	G	G	**C**	A	G	T	C	G
Clade 6	6	1	A	A	G	G	C	G	G	G	G	A	**A**	T	C	G
Clade 7	7	1	A	A	G	G	C	G	G	G	G	**C**	G	T	C	G
Clade 8	8	1	A	A	G	G	C	G	G	G	G	A	G	**A**	C	G
Clade 9	9	16	A	A	G	G	C	G	G	G	G	A	G	T	**A**	G
Clade 10	10	9	A	A	G	G	C	G	G	G	G	A	G	T	C	**A**
new clade	11	6	A	A	G	G	C	G	G	G	G	A	G	T	C	**G**

Total no. 56

* SNP position in the revised *Ma*p K10 sequence (Accession number: AE016958.1).

Defining SNP base marked in bold type

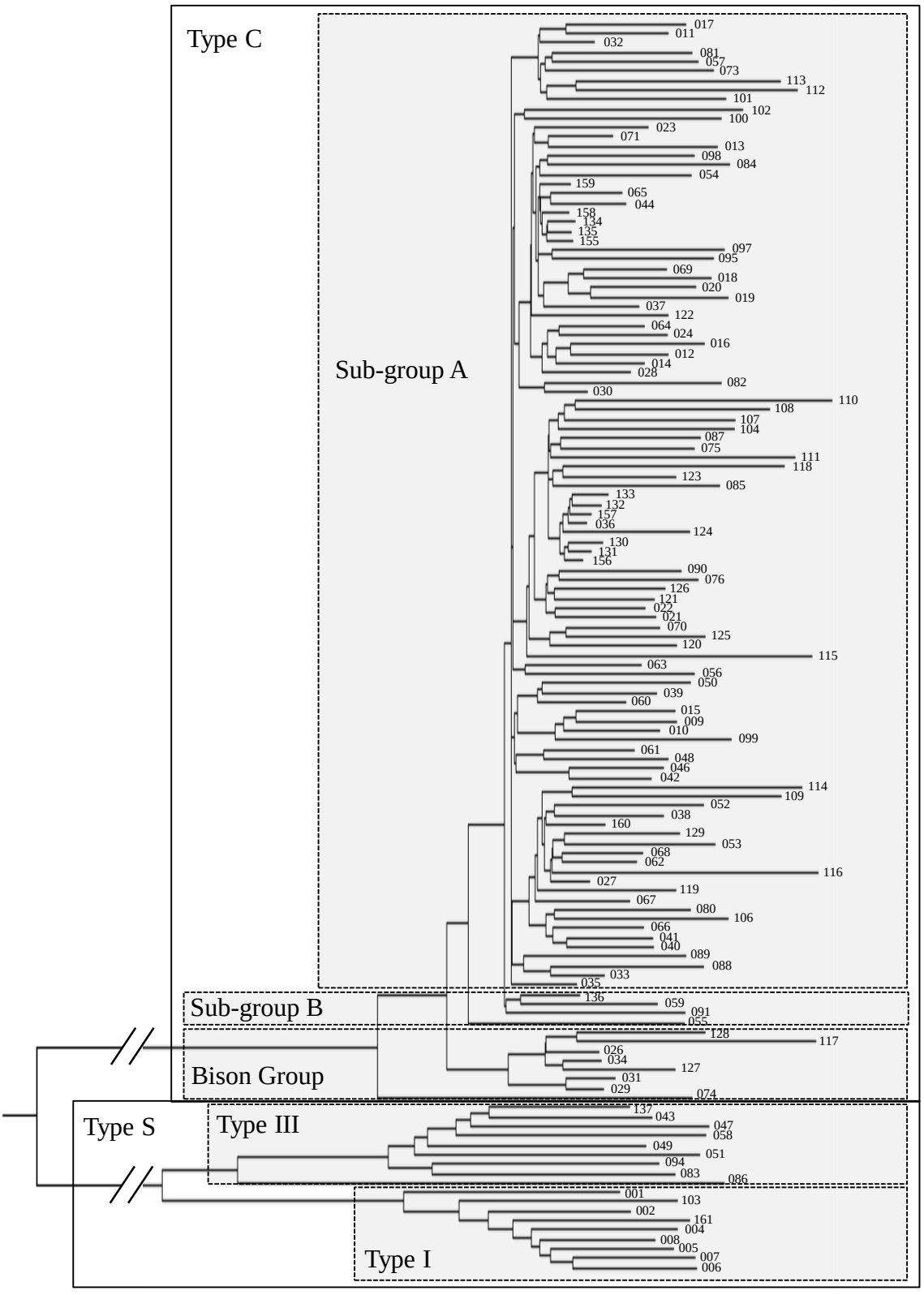
Type C
Sub-group A
Sub-group B
Bison Group
Type S
Type III
Type I
017
011
032
081
057
073
113
112
101
102
100
023
071
013
098
084
054
159
065
044
158
134
135
155
097
095
069
018
020
019
037
122
064
024
016
012
014
028
082
030
110
108
107
104
087
075
111
118
123
085
133
132
157
036
124
130
131
156
090
076
126
121
022
021
070
125
120
115
063
056
050
039
060
015
009
010
099
061
048
046
042
114
109
052
038
160
129
053
068
062
116
027
119
067
080
106
066
041
040
089
088
033
035
136
059
091
055
128
117
026
034
127
031
029
074
137
043
047
058
049
051
094
083
086
001
103
002
161
004
008
005
007
006

Fig. 1

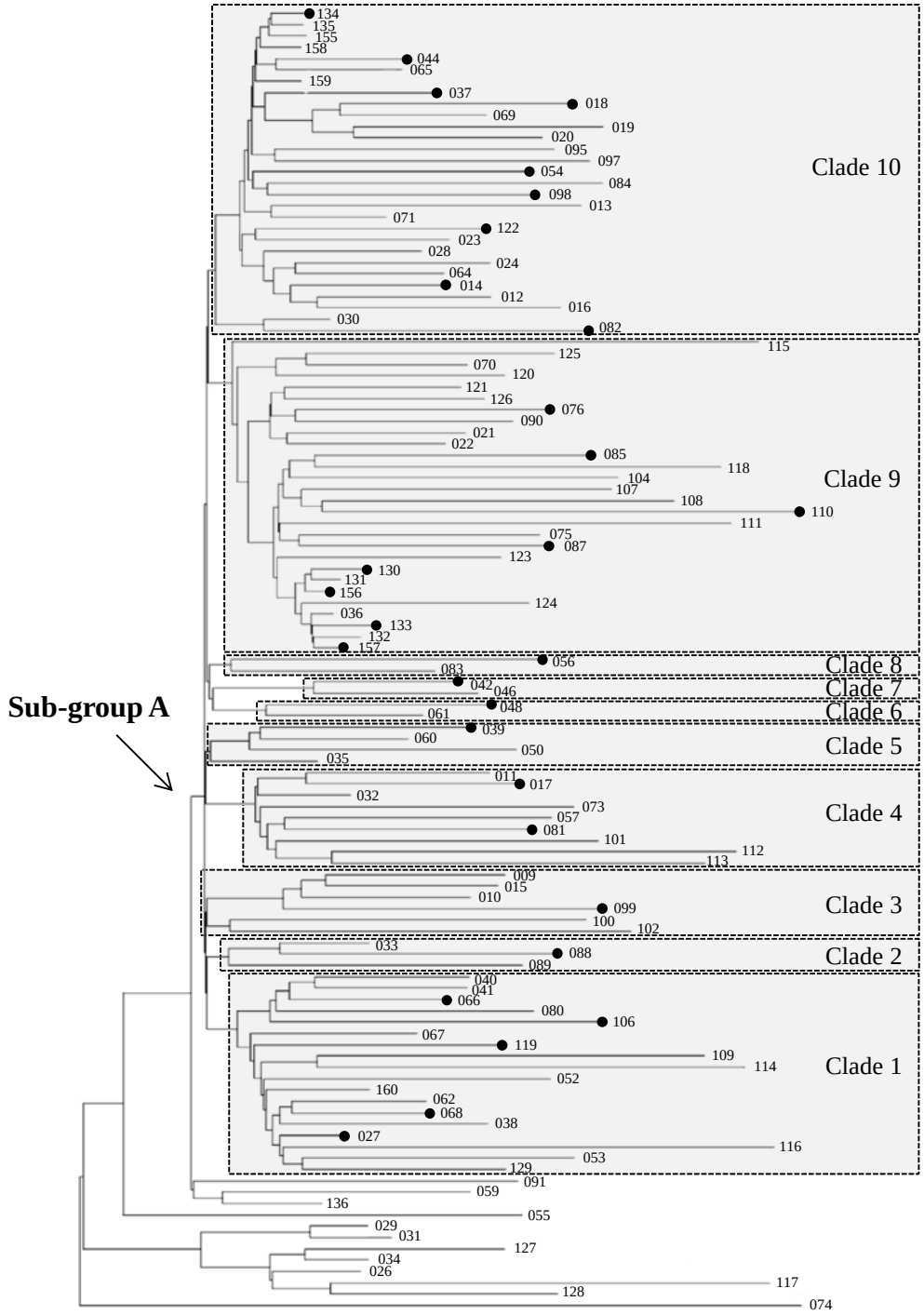
Sub-group A
Clade 10
Clade 9
Clade 8
Clade 7
Clade 6
Clade 5
Clade 4
Clade 3
Clade 2
Clade 1
134
135
155
158
044
065
159
037
018
069
019
020
095
097
054
084
098
013
071
122
023
028
024
064
014
012
016
030
082
115
125
070
120
121
126
076
090
021
022
085
118
104
107
108
110
111
075
087
123
130
131
156
124
036
133
132
157
056
083
042
046
048
061
039
060
050
035
011
017
032
073
057
081
101
112
113
009
015
010
099
100
102
033
088
089
040
041
066
080
106
067
119
109
114
052
160
062
068
038
027
116
053
129
091
059
136
055
029
031
127
034
026
117
128
074

Fig. 2

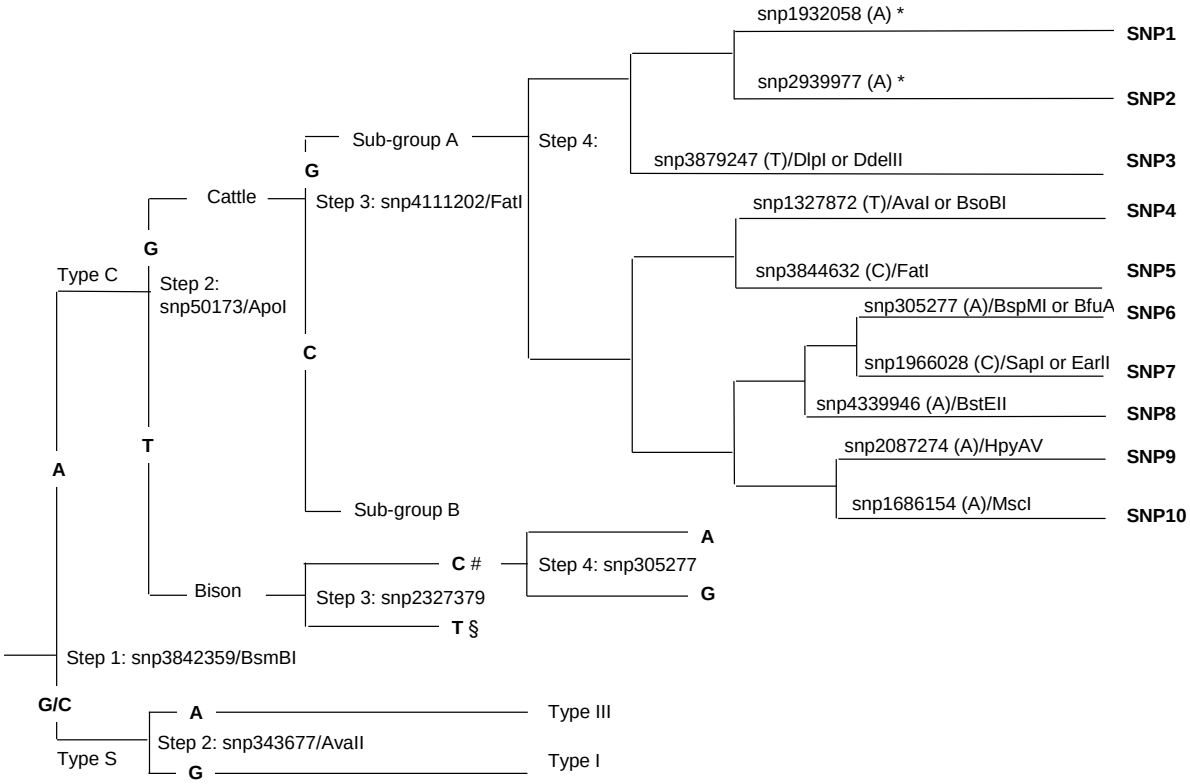
snp1932058 (A) *
SNP1
snp2939977 (A) *
SNP2
Sub-group A
Step 4:
snp3879247 (T)/DlpI or DdelII
SNP3
G
Cattle
Step 3: snp4111202/FatI
snp1327872 (T)/AvaI or BsoBI
SNP4
G
Type C
Step 2:
snp50173/ApoI
snp3844632 (C)/FatI
SNP5
snp305277 (A)/BspMI or BfuA
SNP6
C
snp1966028 (C)/SapI or EarII
SNP7
snp4339946 (A)/BstEII
SNP8
T
snp2087274 (A)/HpyAV
SNP9
A
Sub-group B
snp1686154 (A)/MscI
SNP10
A
C #
Step 4: snp305277
Bison
G
Step 3: snp2327379
T §
Step 1: snp3842359/BsmBI
G/C
A
Type III
Step 2: snp343677/AvaII
Type S
G
Type I

Fig. 3