Phylogenomics picks out the par excellence markers for species phylogeny in the genus *Staphylococcus*.

Lesley Hoyles 22/9/2018 18:33
Deleted: the

Authors

Lucía Graña-Miraglia[1], César Arreguín-Pérez[2], Gamaliel López-Leal[3], Alan Muñoz[1], Ángeles Pérez-Oseguera[1], Estefan Miranda-Miranda[2], Raquel Cossío-Bayúgar[2] and Santiago Castillo-Ramírez[1*].

[1]Programa de Genómica Evolutiva, Centro de Ciencias Génomicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México.

[2]Centro Nacional de Investigación Disciplinaria en Parasitología Veterinaria del Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Carr. Fed. Cuernavaca-Cuautla No. 8534, Jiutepec, Morelos, 62550 México.

[3]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México.

***Correspondence:** Santiago Castillo-Ramírez, iago@ccg.unam.mx

ABSTRACT

Although genome sequencing has become a very promising approach to conduct microbial taxonomy, few labs have the resources to afford this especially when dealing with data sets of hundreds to thousands of isolates. The goal of this study was to identify the most adequate loci for inferring the phylogeny of the species within the genus *Staphylococcus*; with the idea that those who cannot afford whole genome sequencing can use these loci to carry out species assignation confidently. We retrieved 177 orthologous groups (OGs) by using a genome-based phylogeny and an average nucleotide identity analysis. The top 26 OGs showed topologies similar to the species tree and the concatenation of them yielded a topology almost identical to that of the species tree. Furthermore, a phylogeny of just the top 7 OGs could be used for species assignment. We sequenced four staphylococcus isolates to test the 26 OGs and found that these OGs were far superior to commonly used markers for this genus. On the whole, our procedure allowed identification of the most adequate markers for inferring the phylogeny within the genus *Staphylococcus*. We anticipate that this approach will be employed for the identification of the most suitable markers for other bacterial genera and can be very helpful to sort out poorly classified genera.

Lesley Hoyles 22/9/2018 18:33
Deleted: just

Lesley Hoyles 22/9/2018 18:34
Deleted: orthologous groups

Lesley Hoyles 22/9/2018 18:35
Deleted: can

Lesley Hoyles 22/9/2018 18:35
Deleted: to carry out

Lesley Hoyles 22/9/2018 18:35
Deleted: assignation

Lesley Hoyles 22/9/2018 18:35
Deleted: S

Lesley Hoyles 22/9/2018 18:35
Deleted: i

Lesley Hoyles 22/9/2018 18:35
Deleted: yin

Lesley Hoyles 22/9/2018 18:35
Deleted: g

Lesley Hoyles 22/9/2018 18:36
Deleted: not very well

INTRODUCTION

The study of ecology and evolution has been substantially transformed by omics technologies. Remarkably, the use of these technologies has allowed essential questions in evolutionary biology to be addressed and has advanced our knowledge in many biological processes (López-Leal, Tabche et al. 2014, Joseph, Marti et al. 2015, Joseph, Cox et al. 2016). The precise identification of the different species within any given genus is highly valuable for many branches of microbiology. For instance, as far as clinical microbiology is concerned, this is instrumental in establishing which

Lesley Hoyles 22/9/2018 18:38
Deleted: the

Lesley Hoyles 22/9/2018 18:38
Deleted: to address

Lesley Hoyles 22/9/2018 18:38
Deleted: to

species are human/animal pathogens and which are just regular commensal organisms; whereas in microbial ecology species assignation is very helpful in defining the niche-range of the different species (Becker, Margos et al. 2016).

Although whole-genome sequencing is the best method for genotyping bacterial isolates and, therefore, to conduct species assignation/identification, this approach is not yet affordable for routine use in many laboratories all over the world. This is especially true for low- and middle-income countries in which the amount of money invested in science is not as much as in developed countries. For this particular purpose, and from a theoretical point of view, orthologous genes are the perfect candidates for inferring the phylogeny of species, as they should reflect the species tree of the taxa considered. As per definition, orthologous genes are the ideal candidates to track the sequence of past of speciation events within a given lineage (Fitch 2000). Therefore, in order to infer the phylogeny of the species, a clear-cut ascertainment of orthologous genes is of paramount importance and phylogenomic pipelines can be implemented to try to identify the potential orthologous genes. Orthologous relationships are context-sensitive and a gene family could be mainly composed of orthologous genes at one taxonomic level but if one considers a higher taxonomic level many more non-orthologous genes will appear. Furthermore, orthologous relationships could involve one to many relationships (Gabaldón and Koonin 2013). The word co-orthologue was coined to describe that exact situation in which a genome has more than one orthologous gene (Gabaldón and Koonin 2013). Hence, from a practical point of view (i.e. operational implementation), orthologous genes with just one gene per genome among the taxa considered should be the best markers to delineate the history of the species.

Lesley Hoyles 22/9/2018 18:39
Deleted: the

The genus *Staphylococcus* has a few dozens of species of Gram-positive bacteria, which are commensals colonizing the skin and mucous membranes of some mammals and birds. However, some of these species have a clear clinical and economic relevance, as they are a frequent cause of

infection in humans, livestock, and domestic animals. Although the genus is very well known for the human opportunist pathogen *S. aureus*, which is famous worldwide as major source of nosocomial infections (Challagundla, Reyes et al. 2018, Frisch, Castillo-Ramirez et al. 2018), there are some other species such as *S. epidermidis*, *S. lugdunesis*, *S. saprophyticus* and *S. schleiferi* that have been associated with human infections. Several molecular-based methods have been introduced to carry out species identification within the genus *Staphylococcus* (Kwok and Chow 2003, Ghebremedhin, Layer et al. 2008, Sasaki, Tsubakishita et al. 2010, Lamers, Muthukrishnan et al. 2012) . PCR along with sequence analysis of the genes *dnaJ*, *tuf*, *sodA*, *rpoB*, *hsp60* and *nuc* have been used to differentiate *Staphylococcus* species (Kwok and Chow 2003, Ghebremedhin, Layer et al. 2008, Sasaki, Tsubakishita et al. 2010, Lamers, Muthukrishnan et al. 2012) . Although they have been very useful and demonstrate a better resolution than the 16S rRNA gene (Ghebremedhin, Layer et al. 2008), these genes show different amounts of genetic diversity and, therefore, varying levels of discriminatory power for the different species. Thus, depending on the gene and the species, low-level resolution or even mis-identification can occur. Clearly, an ideal solution would be to conduct whole-genome sequencing of all the isolates considered to tell apart the different species. In terms of genome sequences, *Staphylococcus* is one of the few bacterial genera for which many genomes are available.

In this study, we address the question of which genes are the best inter-species markers (*bona fide* orthologous genes) for the genus *Staphylococcus* and, by applying a phylogenomic approach, we provide a list of the top candidates for species assignation within this genus. Underfunded groups could use these top candidates to accurately conduct species assignation without the necessity of using whole-genome sequencing. Furthermore, this approach could be used for many other species/genera to identify the most suitable markers for inferring the phylogeny of the taxa under investigation.

Lesley Hoyles 22/9/2018 18:42
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Deleted: try

Lesley Hoyles 22/9/2018 18:43
Deleted: ing

Lesley Hoyles 22/9/2018 18:43
Deleted: the

Lesley Hoyles 22/9/2018 18:43
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:43
Deleted: s

Lesley Hoyles 22/9/2018 18:44
Deleted: *this*

Lesley Hoyles 22/9/2018 18:44
Formatted: Font:Italic

METHODS

Genomes and homologous groups

We downloaded 265 publically available complete genomes (see Supplementary Table 1). These cover 46 different species from the genus *Staphylococcus* and are a good representation of the host range within this genus. We ran CheckM (Parks, Imelfort et al. 2015) on these genomes to discard poorly sequenced genomes (i.e. with contamination or incomplete) and only one of them (marked in red in Supplementary Table 1) did not pass the criteria (≥95% complete and with ≤5% contamination) and thus were not included in the rest of the analyses. We also sequenced the genomes of 4 bacterial isolates (see Supplementary Table 2) obtained from hemolymph or hypostome exudates from adult ticks showing signs of bacterial infection and reared over experimentally infested bovines in the Centro Nacional de Investigación Disciplinaria en Parasitología Veterinaria (CENID-PAVET-INIFAP), Jiutepec, Morelos, México. Of note the 4 isolates sequenced were confirmed to be *Staphylococcus* spp by biochemical tests and 16S rRNA gene sequence analysis. The isolate INIFAP 005-08 was initially identified as *S. saprophyticus* by its physical and biochemical characteristics (Gram-positive coccus, catalase-positive, novobiocin-resistant, absence of coagulase, gelatinase and caseinase activity). This isolate was also able to produce acids by fermenting glycerol, lactose, D(+) mannose, sucrose and turanose; however, it was unable to use L(+) arabinose or L(+) lactose. This isolates was classified as *S. xylosus* as per 16S rRNA gene analysis. The isolates INIFAP 002-16 and INIFAP 004-15 were identified as Gram-positive cocci, catalase-positive and ribotyped by the 16S rRNA gene, which exhibited positive identity for *S. xylosus*. INIFAP 009-16 was identified as Gram-positive coccus, catalase-positive and identified as *S. succinus* using the 16S rRNA gene. This same isolate was identified as *S. carnosus* using the API 20E (bioMérieux) system. The Nextera XT DNA Library Prep Kit was used for the sequencing libraries and Agilent High Sensitivity DNA Kit was employed for quality control. The isolates were sequenced using an Illumina MiSeq platform, with a 2X250 bp configuration; the genome sequencing was conducted at Instituto Nacional de Medicina Genómica

Lesley Hoyles 22/9/2018 18:44
Deleted: was

Lesley Hoyles 22/9/2018 18:45
Deleted: ,

Lesley Hoyles 22/9/2018 18:45
Deleted: *S*

Lesley Hoyles 22/9/2018 18:45
Deleted: i

Lesley Hoyles 22/9/2018 18:45
Formatted: Font:Italic

Lesley Hoyles 22/9/2018 18:45
Deleted: such as

Lesley Hoyles 22/9/2018 18:46
Deleted:

Lesley Hoyles 22/9/2018 18:46
Deleted:

Lesley Hoyles 22/9/2018 18:47
Deleted: cocci

Lesley Hoyles 22/9/2018 18:46
Deleted: bacteria

Lesley Hoyles 22/9/2018 18:46
Deleted:

Lesley Hoyles 22/9/2018 18:46
Deleted: ;

Lesley Hoyles 22/9/2018 18:46
Deleted:

Lesley Hoyles 22/9/2018 18:47
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:47
Deleted:

Lesley Hoyles 22/9/2018 18:47
Deleted: bacteria

Lesley Hoyles 22/9/2018 18:47
Deleted:

Lesley Hoyles 22/9/2018 18:47
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:47
Deleted:

Lesley Hoyles 22/9/2018 18:47
Deleted: cocci bacteria

Lesley Hoyles 22/9/2018 18:47
Deleted:

Lesley Hoyles 22/9/2018 18:48
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 18:48
Deleted: 2010

(http://www.inmegen.gob.mx) in Mexico City. Prior to assembling the genomes, we employed the SolexQA v.3.7.1 program (Cox, Peterson et al. 2010) to trim the reads. We used Velvet version 1.2.09 (Zerbino and Birney 2008) and Spades v3.11.0 (Bankevich, Nurk et al. 2012) to carry out *de novo* assembly setting the option careful and we tested the following k-mer sizes: 51,61,71,81,91,101,111,121,127. We did not consider contigs smaller than 300 bp in the final assemblies. In all cases, Spades outperformed Velvet in terms of the number of contigs and the N50 statistics. Again, we used CheckM (Parks, Imelfort et al. 2015) to evaluate the quality of the genomes sequenced by us; notably all the 4 isolates were shown to be ≥95% complete and with ≤5% contamination and therefore reliable for downstream analyses. The details of the genomes assemblies are provided in Supplementary Table 2. These seem to be good genome assemblies as judged by the high median coverage and the low number of contigs for each one of the newly sequenced isolates. These 4 genomes have been submitted to the GenBank and have the following accession numbers: PIZQ01000000, PIZN00000000, PIZP00000000, PIZO00000000 (BioProject PRJNA421192). Then, we employed PROKKA v1.11 (we set the genus option to *Staphylococcus*) to annotate all the genomes sequences (newly sequenced and publically available) to have a consistency annotation for all of them. To get the orthologous groups (OGs) with one-to-one relationships among the species, we only considered single gene families (SGF); these were constructed running BLASTP searches with an e-value of 1.0 e^{-30} between the genome of *S. aureus* TW20 and the rest of the genomes. We kept all the cases where there was only one hit per genome and requiring that the seed from TW20 and the hit aligned ≥ 60 % of their lengths and were ≥ 45 % identical - the rationale behind these two criteria was to make sure that we had whole genes and not just domains. Then, for the all the SGF we constructed DNA alignments in frame via the program Fast statistical alignment version 1.5.9 (Bradley, Roberts et al. 2009), setting the option --nucprot that aligns nucleotide sequences taking into account the protein space. We conducted recombination analysis on each of the SGFs using PhiTest (Bruen, Philippe et al. 2006) that was implemented via the PhiPack program, using a window size of 50 bp. Additionally, we also determined the

Lesley Hoyles 22/9/2018 18:49
Deleted: both the

Lesley Hoyles 22/9/2018 18:49
Deleted:

Lesley Hoyles 22/9/2018 18:49
Deleted: ones

Lesley Hoyles 22/9/2018 18:49
Deleted: the

Lesley Hoyles 22/9/2018 18:49
Deleted: ones

Lesley Hoyles 22/9/2018 18:50
Deleted:

nucleotide diversity for each one of the SGFs using R pegas library function nuc.div() with the default parameters. We carried out a function enrichment analysis as follows: first, we conducted a Gene Onthology (GO) annotation via InterProScan version 5 using the genes from the *S. aureus* TW20 genome as a reference. Then, we conducted a GO enrichment analysis via Blast2GO PRO using a Fisher's Exact Test (https://www.blast2go.com/), the analysis was conducted at three different GO levels: Molecular Function (MF), Biological process (BP) and Cellular Component (CC). Then to evaluate the results, we used a False Discovery Rate of 0.05 and the Benjamini-Hochberg correction was used to account for multiple testing.

Phylogenetic reconstructions, average nucleotide identity analysis and neighbour nets.

We performed phylogenetic reconstructions for every SGF that did not show recombination signals as per PhiTest (Bruen, Philippe et al. 2006) and for the super alignment (read below). All the gene trees were constructed with RaxML version 8.2.11 (Stamatakis 2014) executing 10 inferences on the alignment, using 10 distinct randomized Maximum Parsimony (MP) trees and with the GTR+G+I model. Both the Shimodaira-Hasegawa topology test and the Robinson-Foulds distance were also conducted via RAxML with the default settings. It is known that species tree estimation is not a trivial matter and we employed a previous approach that gave trustable results (Castillo-Ramirez and Gonzalez 2008). We created a super alignment concatenating all the SGFs that did not have signals for recombination and on this alignment a ML phylogeny was constructed also through RAxML, this time executing 20 independent inferences starting from 20 different MP trees and with GTR+G+I model. For this ML phylogeny bootstrap replicates were generated again employing RaxML, using the -x option that implements a fast algorithm for bootstrapping. The average nucleotide identity (ANI) analysis was run via the python module pyani (http://widdowquinn.github.io/pyani/), specifically we employed the ANIm method (Richter and Rosselló-Móra 2009). To determine the GC content and the proportion of variable sites for each SGF, we used the function summary from the program AMAS (Borowiec 2016). In order to

Lesley Hoyles 22/9/2018 18:51
Deleted: h

visualize the conflicting phylogenetic signals in the genes *rpoB* and *tuf* we used SplitTree4 (Huson and Bryant 2005) to construct Neighbour nets with uncorrected P distances.

RESULTS

Defining the species tree, confirming genome affiliations and the set of orthologous genes

We focused on the genus *Staphylococcus* as it has clear clinical and veterinary relevance and, due to that, it has been extensively covered in terms of genome sequences. The data set employed for this study represents 46 species, incorporating 40 type strains, and included a total of 269 genomes for the analyses (see Supplementary Tables 1 and 2). Notably, this data set has species with a wide host-range covering many of the niches described for this genus. First, we determined the SGFs, as these are potential candidates to be orthologous genes with one-to-one relationships, and found 208 SGFs. However, recombination could have affected them and, therefore, we conducted recombination tests on them and around 15% of them (31 SGF) showed signals of recombination and were discarded, which left 177 SGFs as good candidates to be OGs; the list of these 177 SGFs is provided in Supplementary Table 3. We employed a previously used strategy [20] to approximate the species tree. This is the total evidence approach, in which all the 177 SGFs without signals of recombination are concatenated and treated as if they were a single marker on which a Maximum Likelihood (ML) phylogeny was constructed (see Figure 1) – this ML phylogeny was our Proxy for the Species Tree Topology (PSTT). From Figure 1 one can see that type strains are scattered throughout the tree and that most of the isolates from individual species tend to form monophyletic groups and these groups are very well-supported as most of them have bootstrap values higher than 80 (see orange dots in the phylogeny). However, we also noted a region on the tree (shaded areas), where different species intermingle together. This is caused by two species, namely *S. warneri* and *S. saccharolyticus,* which do not form monophyletic groups and clearly some strains from these species have been mislabelled. Additionally, as an independent strategy, we also carried out an ANI analysis to calculate the relatedness of the strains to the type strains included in this study (see

Lesley Hoyles 22/9/2018 18:52
Deleted: this

Lesley Hoyles 22/9/2018 18:52
Formatted: Font:Italic

Lesley Hoyles 22/9/2018 18:53
Deleted: single gene families (

Lesley Hoyles 22/9/2018 18:53
Deleted:)

Lesley Hoyles 22/9/2018 18:53
Deleted: orthologous groups (

Lesley Hoyles 22/9/2018 18:53
Deleted:)

Lesley Hoyles 22/9/2018 18:54
Comment [1]: Format of references should be consistent throughout the manuscript.

Lesley Hoyles 22/9/2018 18:57
Deleted: proxy

Lesley Hoyles 22/9/2018 18:57
Deleted: species

Lesley Hoyles 22/9/2018 18:57
Deleted: tree

Lesley Hoyles 22/9/2018 18:57
Deleted: topology

Lesley Hoyles 22/9/2018 19:13
Deleted: average nucleotide identity

Figure 2). This analysis shows that in terms of their taxonomy most of the genomes have been properly labelled. For instance, all the strains designated as *S. lugdunensis* clustered with the type strain, *S. lugdunensis* NCTC 12217^{T}, with identity percentages well above 95%. The same applies to most of the other species; see for example *S. gallinarum* or *S. simulans* where genomes labelled as belonging to each of these species clustered tightly (again above the 95%) with the type strains. On the other hand, the mis-classification previously noted in the phylogeny is also evident in this analysis, as the strains of both *S. warneri* and *S. saccharolyticus* do not cluster all together in each case and their identity percentages are well below 95% (see green labels in Figure 2). We then carried out a GO enrichment analysis to have an idea of the overrepresented functions of the 177 SGFs (see Supplementary Table 4), as expected many of the enriched biological processes and molecular functions have to do with housekeeping functions; the top 2 GO molecular functions were ATP binding and GTP binding, whereas the top 2 Biological processes were DNA repair and fatty acid biosynthesis (see Supplementary Table 4 for more details). Taken together, these results demonstrate that the 177 SGFs seem to be real OGs and that most of the genomes have a proper affiliation regarding their taxonomy.

Ranking the orthologous groups.

Then we analysed which of the 177 SGFs, from now on OGs, could be the best markers to infer the evolutionary relationships among the species. For that end, we established how similar each one of the single gene trees (from the 177 OGs) was compared to the PSTT; this was done employing the Robinson and Foulds (RF) distance between the single gene trees and the PSTT – this distance gives the number of bipartitions that are not shared by the two trees under consideration. We also used π, a common measure of genetic diversity, to establish the amount of genetic variation for each OGs. Figure 3 gives the percentage of similarity of the 177 OG trees to the PSTT and the nucleotide diversity for each of the 177 OGs; this figure shows that no single OG yielded the same topology as the PSTT. Furthermore, we also noted that every single OG has its own topology – not shared by

Lesley Hoyles 22/9/2018 18:56
Comment [2]: Include reference to Chun et al. (2018) here re species delineation based on ANI >95-96 %.

Lesley Hoyles 22/9/2018 18:55
Formatted: Superscript

Lesley Hoyles 22/9/2018 18:56
Deleted: s

Lesley Hoyles 22/9/2018 18:56
Deleted: Gene Ontology (

Lesley Hoyles 22/9/2018 18:56
Deleted:)

any other OG – as none of all the pairwise comparisons of the OGs trees gave a RF equal to 0 (a RF distance of 0 implies that the two topologies in comparison are the same, see Supplementary Figure 2). This Figure also shows that most of the OGs have nucleotide diversity values higher than 0.15, which make them good candidates for phylogenetic markers. Of note, for two species (*S. aureus* and *S. epidermidis*) we also computed the intra-species diversity and it seems that these OGs even at this level have a good amount of genetic diversity (see Supplementary Figure 3). We found that the top 26 OGs were similar to the PSTT, as all of them have RF distances below 214, indicating that they are more than 60 % identical to the PSTT. Notably these 26 OGs present good values of nucleotide diversity (all but two showing values higher than 0.2 and the average for the 26 being 0.26), which is very convenient for their use as to phylogenetic markers. Table 1 provides details about these OGs such as RF distance PSTT, nucleotide diversity, function and the ID in the *S. aureus* TW20 genome. Remarkably, when we concatenated these 26 OGs and constructed a phylogeny, this tree showed a topology pretty similar to the PSTT (see Figure 4, panel A), the RF distance was 74 being 86.1% identical to the PSTT. Furthermore, using only the best 7 OGs (the bold candidates from Table 1), we got a similar result – that is the concatenated ML phylogeny of these 7 OGs is almost 80% identical to the PSTT (see panel B, Figure 4). Here it is worth mentioning that although these two phylogenies (Figure 4) are not 100% identical to the PSTT, the two phylogenies recovered almost all the species as monophyletic groups with very good bootstrap values (> 80%), see blue dots in both phylogenies - the only exceptions being the miss-classifications noted also in the PSTT and ANI analysis. In considering this part, although no single gene tree of the 177 OGs yielded the PSTT, just using the best 26 (and even just the top 7) OGs we were able to recover a topology pretty similar to the PSTT.

The inadequacy of usual markers and the soundness of our strategy

Then, we wanted to know how a set of commonly used markers for inferring the phylogenetic relationships within this genus compared to our list. We chose the genes *dnaJ, tuf, sodA, rpoB* and

Lesley Hoyles 22/9/2018 18:58
Deleted: —

hsp60 as these are the set of genes most used in previous studies [12,25]. Most of them had clear issues to be considered suitable markers given the criteria we employed to define our genuine OGs. Notably two of them, *tuf* and *rpoB*, did have signals for recombination as shown by their Neighbour nets and the PhiTest (see Supplementary Figure 1) and their nucleotide diversity was very low (below 0.15, see Figure 3 orange dots). On the other hand, *sodA* and *hsp60* were not single gene families, as they did have more than one gene per genome in some species. Among these usual markers, only *dnaJ* seems to be a good candidate in as much as it is a SGF, did not have signals of recombination and has a nucleotide diversity value above 0.2 (see Figure 3). The shortcoming of most of these genes is not totally unexpected, as these genes were not selected based on phylogenetic criteria; however, it is clear that most of these genes do not seem to be a good option to infer the history of the species.

Finally, to prove the validity of our phylogenomic approach, in the data set here employed we included the genome sequences of four bacterial isolates collected by us from cattle ticks and phenotypically classified as members of the genus *Staphylococcus* (see Supplementary Table 2). Initially these isolates were classified using the 16S rRNA gene sequence (see Supplementary Table 2); however, in 2 of these cases, namely INIFAP 009-16 and INIFAP 002-15, the initial classification was incorrect. For instance, INIFAP 009-16 was classified as *S. succinus* by its 16S rRNA gene sequence but the PSTT placed this isolate with *S. xylosus*. In the case of INIFAP 002-15, whereas PSTT assigned it to *S. succinus,* the starting classification via the 16S rRNA gene was *S. xylosus*. Importantly, in the concatenate alignments using our 26 OGs (Table 1) the isolates were properly classified (see panel A, Figure 4). Furthermore, the ML phylogeny based on the concatenated alignment of the top 7 OGs also recovers the true affiliation of these isolates (see panel B, Figure 4). To sum up, our genuine OGs are a much better option than the usual markers to establish the phylogeny of the species and even just a few of these (the top 7) are able to adequately carry out taxonomy classification of the newly sequenced bacterial isolates.

Lesley Hoyles 22/9/2018 19:10
Comment [3]: Consistency of reference formatting?

Lesley Hoyles 22/9/2018 19:11
Deleted: - these isolates were sequenced using an Illumina MiSeq platform (see methods)

Lesley Hoyles 22/9/2018 19:11
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 19:11
Formatted: Font:Not Italic

Lesley Hoyles 22/9/2018 19:11
Formatted: Font:Not Italic

DISCUSSION

The main goal of our study was to establish the best markers for species identification in the genus *Staphylococcus.* To try to be as comprehensive as possible, we used a data set with over 45 species and a total of 269 genomes from this genus that broadly represents most if not all the niches covered by this genus. Here we identified 177 OGs and ranked them according to their potential as phylogenetic markers; clearly, this set of markers should be useful not only for ecological and evolutionary studies but also for clinical and biotechnological purposes. In addition, our phylogenomic approach allowed us to identify two species (*S. warneri* and *S. saccharolyticus*) in which mislabelling of deposited genomes has occurred. This issue of misclassification is not that rare, as genome sequences submitted in public databases many times do not undergo a proper inspection in terms of microbial taxonomy. However, the combination of these phylogenomic approaches (genome-based phylogeny and ANI analysis) can be very useful to pinpoint those cases of misclassification.

We want to emphasise that our study has two major contributions: one is the list of adequate markers for inferring the phylogeny of the species within the genus *Staphylococcus* and the other is the phylogenomic strategy that we used to identify such markers. Regarding the first major contribution, given that the average gene content of the species here analysed is 2413, our 177 OGs represent barely the 7 % of the genome of these species. However, we need to emphasise here that, for merely practical reasons, we focused on orthologous genes with one-to-one relationships and, very likely, there are many more orthologous genes within these species that did not pass our criteria given that they are not SGF. Despite the fact that all these OGs are potential good markers, in as much as their histories do not deviate significantly from the history of the species, they do differ among them in the amount of phylogenetic signal they each present. Importantly, no single OG was able to exactly represent the PSTT and no two OGs yielded the same topology; clearly, we

Lesley Hoyles 22/9/2018 19:12
Deleted: y

Lesley Hoyles 22/9/2018 19:14
Deleted: this

Lesley Hoyles 22/9/2018 19:14
Formatted: Font:Italic

Lesley Hoyles 22/9/2018 19:14
Deleted:

Lesley Hoyles 22/9/2018 19:14
Deleted:

found a cloud of topologies. This is in agreement with a previous study that found that only one of the several hundreds of OGs reflected the species tree (Castillo-Ramirez and Gonzalez 2008), although this focused on different evolutionary scales and different bacteria. It is worth mentioning that the top 26 OGs have good values of nucleotide diversity (all but two have values higher than 0.21), and are localized in different parts of the chromosome and present different functions. We think these 26 OGs represent a really good set of markers for inferring the PSTT; actually concatenating the best 7 it was possible to get almost exact same topology as the PSTT. Here, we want to highlight that using the phylogeny of the top 7 OGs all but the 2 species with issues of mis-classification were recovered as monophyletic groups and thus it seems that just using these 7 markers species assignation can be carried out. Clearly, this set of markers could be very useful for underfunded laboratories working with *Staphylococcus* isolates. Although from a genotyping perspective, ideally one would want to use whole-genome sequencing for species identification, many laboratories in the world (specially in developing countries) still cannot afford the sequencing of tens to hundreds of isolates. Thus, this short list of markers should be extremely useful for those who only can sequence a few loci, which is often the case for clinical, environmental and evolutionary biology microbiologists in developing countries. Clearly, the markers highlighted in this study are a much better option that the commonly used markers, as the latter seem to exhibit obvious flaws (more than one copy per genome, signals of recombination, low nucleotide diversity values) for inferring the history of the species. Our strategy, and in turn the list of genes found, proved to be factually sound, even when using just a few markers for the taxonomic assignment of newly sequenced bacterial isolates.

Maybe more important than the first major contribution is the fact that our phylogenomic approach will allow the identification of adequate markers in many different genera – not only from bacteria but also from archaea. This is very important, as gene families showing only orthologous relationships very likely do not extend all over the tree of life – or just for a few gene families.

Lesley Hoyles 22/9/2018 19:15
Deleted: ,

Lesley Hoyles 22/9/2018 19:16
Deleted: ,

Lesley Hoyles 22/9/2018 19:16
Deleted: s

Lesley Hoyles 22/9/2018 19:16
Deleted: , which are

Lesley Hoyles 22/9/2018 19:16
Deleted: ,

Lesley Hoyles 22/9/2018 19:16
Deleted:

Lesley Hoyles 22/9/2018 19:17
Deleted: ,

Lesley Hoyles 22/9/2018 19:17
Deleted: to

Lesley Hoyles 22/9/2018 19:17
Deleted: y

Lesley Hoyles 22/9/2018 19:17
Deleted: extent

Along these lines, it is very likely that many of the markers found here will not work for other genera (i.e. they will not show one-to-one orthologous relationships) as the homologous genes within those genera might have been affected by horizontal gene transfer or duplication and differential loss, or some other molecular event that prevent the history of the gene to reflect the speciation events. We want to highlight that the biology of the species under consideration could have a very important effect on the number of orthologous genes found. For instance, highly recombinogenic species, such as *Neisseria gonorrhoeae* (Ezewudo, Joseph et al. 2015) and *Acinetobacter baumannii* (Grana-Miraglia, Lozano et al. 2017), would have considerably fewer orthologous genes than more clonal species. Nonetheless, the strategy employed by us should work even in these species, although the number of potential orthologous genes should be much less.

Lesley Hoyles 22/9/2018 19:18
Deleted: HGT

Lesley Hoyles 22/9/2018 19:18
Deleted: less

CONCLUSIONS

In summary, here we devised and applied a phylogenomic approach that allowed us to define the most suitable markers for inferring the species phylogeny within the genus *Staphylococcus*. We acknowledge that the effectiveness of these markers for other bacterial genera remains open to investigation. However, we are confident that the phylogenomic approach here implemented can be employed to identify the most suitable markers for other genera. On a broader level, this study has very practical implications, as it provides a general framework for mapping out the best markers to construct the phylogeny of the taxa considered not only at the genus level but also at other taxonomic levels.

Acknowledgments

We are grateful to the editor and the reviewers, as their suggestions have drastically improved our manuscript.

REFERENCES

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham and A. D. Prjibelski (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of computational biology **19**(5): 455-477.

Becker, N. S., G. Margos, H. Blum, S. Krebs, A. Graf, R. S. Lane, S. Castillo-Ramírez, A. Sing and V. Fingerle (2016). "Recurrent evolution of host and vector association in bacteria of the Borrelia burgdorferi sensu lato species complex." BMC genomics **17**(1): 734.

Borowiec, M. L. (2016). "AMAS: a fast tool for alignment manipulation and computing of summary statistics." PeerJ **4**: e1660.

Bradley, R. K., A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes and L. Pachter (2009). "Fast statistical alignment." PLoS computational biology **5**(5): e1000392.

Bruen, T. C., H. Philippe and D. Bryant (2006). "A simple and robust statistical test for detecting the presence of recombination." Genetics **172**(4): 2665-2681.

Castillo-Ramirez, S. and V. Gonzalez (2008). "Factors affecting the concordance between orthologous gene trees and species tree in bacteria." BMC Evol Biol **8**: 300.

Challagundla, L., J. Reyes, I. Rafiqullah, D. O. Sordelli, G. Echaniz-Aviles, M. E. Velazquez-Meza, S. Castillo-Ramírez, N. Fittipaldi, M. Feldgarden and S. Chapman (2018). "Phylogenomic Classification and the Evolution of Clonal Complex 5 Methicillin-Resistant Staphylococcus aureus in the Western Hemisphere." Frontiers in Microbiology **9**: 1901.

Cox, M. P., D. A. Peterson and P. J. Biggs (2010). "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data." BMC bioinformatics **11**(1): 485.

Ezewudo, M. N., S. J. Joseph, S. Castillo-Ramirez, D. Dean, C. Del Rio, X. Didelot, J.-A. Dillon, R. F. Selden, W. M. Shafer and R. S. Turingan (2015). "Population structure of Neisseria gonorrhoeae based on whole genome data and its relationship with antibiotic resistance." PeerJ **3**: e806.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." Trends Genet **16**(5): 227-231.

Frisch, M., S. Castillo-Ramirez, R. Petit, M. Farley, S. Ray, V. Albrecht, B. Limbago, J. Hernandez, I. See and S. Satola (2018). "Invasive methicillin-resistant Staphylococcus aureus USA500 strains from the US Emerging Infections Program constitute three geographically distinct lineages." mSphere **3**(3): e00571-00517.

Gabaldón, T. and E. V. Koonin (2013). "Functional and evolutionary implications of gene orthology." Nature Reviews Genetics **14**(5): 360.

Ghebremedhin, B., F. Layer, W. König and B. König (2008). "Genetic classification and distinguishing of Staphylococcus species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf gene sequences." Journal of clinical microbiology **46**(3): 1019-1025.
Grana-Miraglia, L., L. F. Lozano, C. Velazquez, P. Volkow-Fernandez, A. Perez-Oseguera, M. A. Cevallos and S. Castillo-Ramirez (2017). "Rapid Gene Turnover as a Significant Source of Genetic Variation in a Recently Seeded Population of a Healthcare-Associated Pathogen." Front Microbiol **8**: 1817.
Huson, D. H. and D. Bryant (2005). "Application of phylogenetic networks in evolutionary studies." Molecular biology and evolution **23**(2): 254-267.
Joseph, S. J., D. Cox, B. Wolff, S. S. Morrison, N. A. Kozak-Muiznieks, M. Frace, X. Didelot, S. Castillo-Ramirez, J. Winchell and T. D. Read (2016). "Dynamics of genome change among Legionella species." Scientific reports **6**: 33442.
Joseph, S. J., H. Marti, X. Didelot, S. Castillo-Ramirez, T. D. Read and D. Dean (2015). "Chlamydiaceae genomics reveals interspecies admixture and the recent evolution of Chlamydia abortus infecting lower mammalian species and humans." Genome biology and evolution **7**(11): 3070-3084.
Kwok, A. Y. and A. W. Chow (2003). "Phylogenetic study of Staphylococcus and Macrococcus species based on partial hsp60 gene sequences." International journal of systematic and evolutionary microbiology **53**(1): 87-92.
Lamers, R. P., G. Muthukrishnan, T. A. Castoe, S. Tafur, A. M. Cole and C. L. Parkinson (2012). "Phylogenetic relationships among Staphylococcus species and refinement of cluster groups based on multilocus data." BMC evolutionary biology **12**(1): 171.
López-Leal, G., M. L. Tabche, S. Castillo-Ramírez, A. Mendoza-Vargas, M. A. Ramírez-Romero and G. Dávila (2014). "RNA-Seq analysis of the multipartite genome of Rhizobium etli CE3 shows different replicon contributions under heat and saline shock." BMC genomics **15**(1): 770.
Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." Genome research: gr. 186072.186114.
Richter, M. and R. Rosselló-Móra (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proceedings of the National Academy of Sciences **106**(45): 19126-19131.
Sasaki, T., S. Tsubakishita, Y. Tanaka, A. Sakusabe, M. Ohtsuka, S. Hirotaki, T. Kawakami, T. Fukata and K. Hiramatsu (2010). "Multiplex-PCR method for species identification of coagulase-positive staphylococci." Journal of clinical microbiology **48**(3): 765-769.
Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics **30**(9): 1312-1313.
Zerbino, D. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research: gr. 074492.074107.