PRACTICALLY DELINEATING BACTERIAL SPECIES WITH GENEALOGICAL CONCORDANCE

Stephanus N. Venter*, Marike Palmer, Chrizelle W. Beukes, Wai-Yin Chan, Giyoon Shin, Elritha van Zyl, Tarren Seale, Teresa A. Coutinho and Emma T. Steenkamp

Department of Microbiology and Plant Pathology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

Corresponding author*:

Email: fanus.venter@up.ac.za

Tel: +27 12 4204100

Fax: +27 12 4203266

Abstract

Bacterial species are commonly defined by applying a set of predetermined criteria, including DNA-DNA hybridization (DDH) values, 16S rRNA sequence similarity, phenotypic data as well as genome-based criteria such as average nucleotide identity (ANI) or genome-to-genome distance hybridization (GGDH). These criteria mostly allow for the delimitation of taxa that resemble typical bacterial species. Their application is often complicated when the objective is to delineate new species that are characterized by significant population-level diversity or recent speciation. However, we believe that these complexities and limitations can be easily circumvented by recognizing that bacterial species represent unique and exclusive assemblages of diversity. Within such a framework, methods that account for the population processes involved in species evolution are used to infer species boundaries. A method such as genealogical concordance analysis is well suited to delineate a putative species. The existence of the new taxon is then interrogated using an array of traditional and genome-based characters. By making use of taxa in the genera *Pantoea*, *Paraburkholderia* and *Escherichia* we demonstrate in a step-wise process how genealogical concordance can be used to delimit a bacterial species. Genetic, phenotypic and biological criteria were used to provide independent lines of evidence for the existence of that taxon. This approach to species recognition and description is straightforward and applicable to bacterial species especially in the post-genomic era, with increased availability of whole genome sequences. In fact, our results

indicated that a combined genome-based comparative and evolutionary approach would be the preferred alternative for delineating coherent bacterial taxa.

Keywords

Bacterial taxonomy; *Escherichia coli;* Genealogical concordance; *Paraburkholderia*; *Pantoea*; Species recognition

Introduction

Bacterial species provide biologists with a framework to describe, organise and investigate bacterial diversity. Without this framework it would be extremely difficult to understand biological systems and the specific roles and interactions of different bacteria in these systems. However, the species definitions and concepts proposed for bacteria are varied as they are based on different taxonomic, evolutionary and ecological perspectives (Reviewed by Rossello-Mora and Amann 2001; 2015). Although taxonomists generally regard the concept of species as an artificial or man-made idea (Rosselló-Mora and López-López 2008), most accept the possibility that bacterial species could be real units representative of the taxa occuring in nature. For example, following de Queiroz's (2005) view that "*species are separately evolving metapopulation lineages*", Achtman and Wagner (2008) concluded that bacteria "*that form distinct groups owing to cohesive forces are metapopulation lineages and thus form species*".

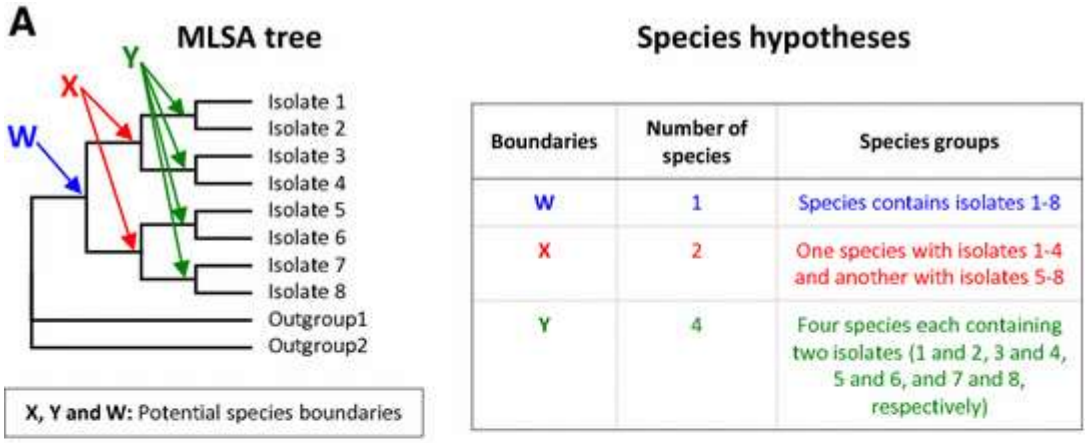
Bacterial systematics currently has a strong focus on coherence within bacterial species. This is illustrated by Rosselló-Móra and Amann (2015), who defined these taxa as "*monophyletic and genomically and phenotypically coherent populations of individuals that can be clearly discriminated from other such entities*". This emphasis on genomic and phenotypic coherence is reminiscent of Mallet's (2001) idea of species as "*multilocus genotypic clusters*". Accordingly, new bacterial species are typically recognized and described based on the collection and integration of a wide range of phenotypic and genotypic data (Tindall et al. 2010). This approach is widely referred to as polyphasic taxonomy (Colwell 1970; Vandamme et al. 1996), which aims to arrange organisms into groups based on the "consensus" of the data collected. Decisions on phylogenetic coherence and monophyly are usually based on 16S rRNA phylogenies (Stackebrandt et al. 2002). A bacterial species is, therefore, recognised by the fact that it has a common origin and possesses a shared set of distinct genetic and phenotypic characteristics.

Despite the wide application of polyphasic taxonomy in bacterial systematics (Rossello-Mora and Amann 2015), the delineation of species is often not straightforward. This is because no clear and logical evolution-based guideline is available for identifying the species boundary. Therefore, when confronted with a diverse set of varying traits and characters, the decision of where to position the limit of what constitutes a bacterial species remains subjective. To increase the objectivity of taxonomic decisions, a pragmatic approach has been to apply empirically-derived quantitative cut-off values for some of the commonly employed parameters used to circumscribe species. Traditionally, these included DNA-DNA reassociation values of ≤70% and 16S rRNA gene sequence similarity values of ≤97% for defining isolates belonging to different species (Wayne et al. 1987; Tindall et al. 2010) With the wide availability of whole genome sequence information, average nucleotide identity (ANI) is also increasingly used where conspecific isolates are characterized by ANI values of ≥95% (Goris et al. 2007; Richter and Rossello-Mora 2009).

The use of a single set of quantitative criteria for delineating bacterial species has been criticised as it assumes that the evolution of all bacterial species is uniform. Speciation is a continuous process and bacterial species do not necessarily evolve at the same rate or are at the same level of divergence (Retchless and Lawrence 2007) . The limits of what constitutes a species are dependent on the combined effects of the processes, mechanisms and biological drivers involved in the evolution of that particular taxon. Therefore, the expectation that predefined cut-off values for any set of biological metrics will allow for the delineation of all species taxa is both naïve and unattainable (Hey 2001; Ereshefsky 2011; Booth et al. 2016).

Multilocus sequence analysis (MLSA) has been proposed as an evolution-based approach to objectively investigate the boundaries between species (Gevers et al. 2005; Brady et al. 2008; Glaeser and Kampfer 2015). Although the resulting phylogenetic trees, based on concatenated gene sequence data, typically resolve the relationships among taxa relatively well, the recognition of species is often still arbitrary and problematic as several well-defined monophyletic groups, each representing an alternative species hypothesis, could be demarcated based on the same MLSA tree (Figure 1A). Based on where the species boundary are placed, the groups on an MLSA tree could either represent distinct species or sub-populations within a single species (Gevers et al. 2005). To overcome this problem, researchers have again proposed the use of MLSA sequence divergence thresholds (Naser et al. 2007) or cut-off values for certain bacterial groups (Vanlaere et al. 2009;

Vandamme et al. 2013). However, such thresholds and cut-offs suffer from the same limitations as those based on DNA-DNA hybridization, 16S rRNA gene sequences and ANI in being subjective and arbitrary. Indeed, Gevers et al. (2005) suggested that the separation of species using MLSA should be guided by additional ecological and genomic data as markedly different processes might have driven their evolution.



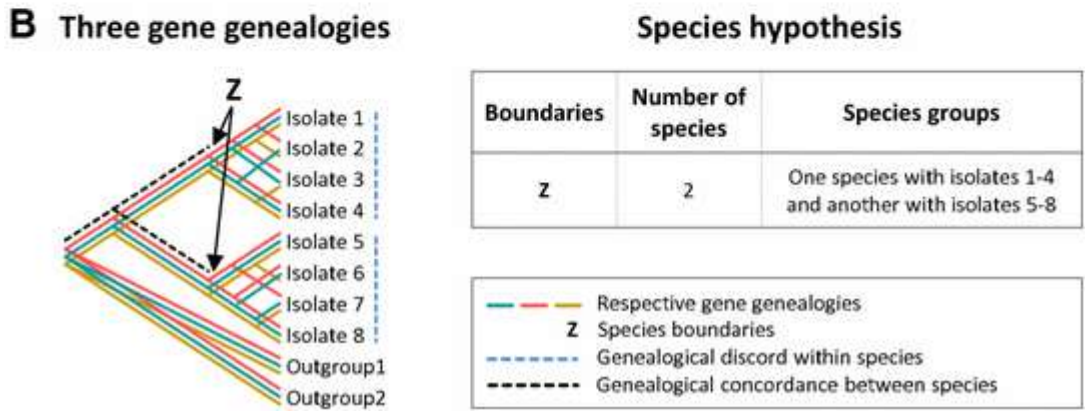


Figure 1. Generation of "species hypotheses" using MLSA (A) and genealogical concordance (B). From an MLSA tree, which is inferred from the concatenated sequence information for several genes, multiple well supported groups of individuals are typically recovered. In a systematics study, these groups depict the various species hypotheses requiring evaluation. Compared to the multitude of possible hypotheses deducible from an MLSA tree, a single species hypothesis typically emerges from a genealogical concordance study. For the latter, genealogies of several independent loci are examined for the transition from branch concordance to discordance, where this transitional point allows delineation of a plausible species hypotheses.

Clearly, one of the prevailing problems, given the current polyphasic taxonomy framework, is the identification of groups that potentially represent distinct species. In this paper we demonstrate how genealogical concordance can be used to objectively identify plausible species hypotheses. For this purpose, the theory behind

genealogical concordance is briefly reviewed. We accordingly provide a proposal (in the form of an easy-to-follow and straightforward workflow) of how genealogical concordance is best incorporated into the current polyphasic taxonomy framework. We believe that this will provide a solid theoretical basis for the interpretation of data typically used during the polyphasic approach. The application and value of this approach are then illustrated by making use of examples from the genera *Pantoea*, *Paraburkholderia* and *Escherichia*.

Genealogical concordance – a convenient and objective way to generate species hypotheses

Genealogical concordance is classified amongst the myriad of phylogenetic species concepts that have been developed since the 1970s (Coyne and Orr 2004). The use of this approach for recognizing species is based on the genealogical decent of a species' genome (Sites and Marshall 2004), where the genotypes and phenotypes of the members of a species are genealogically related (Avise and Ball 1990; Baum and Shaw 1995). Reticulation at the population level (due to within-species genetic processes such as recombination among individuals and inter-population gene flow) typically causes the genealogies inferred from the independent phenotypes and genotypes of individuals to be discordant (Avise and Ball 1990; Rising and Avise 1993; Baum and Shaw 1995). However, given sufficient phylogenetic time (during which lineages diverge due to the cessation of reticulation between lineages and the fixation of character states), the genealogies of these phenotypes and genotypes will become concordant with one another within the species. In the words of Wilson and Brown (1953), who first conceived of the idea of genealogical concordance, "*complete concordance of several known independent characters […] may be a good indication that the population has attained species level*".

With the advances in molecular biology and sequencing technologies, bacterial systematists increasingly have access to large volumes of data for delineating species. In response to such an emphasis on DNA sequence information, Baum and Shaw (1995) made a detailed analysis of how sequence information could be used to delineate species using genealogical concordance. They highlighted two basic assumptions. The first assumption is that the species in question should represent basal taxa (i.e., one species should not contain another species) (Sites and Marshall 2004). In other words, the members of a species should be more closely related to each other than to any organism outside the group. The second assumption is that the species boundary resides at the interface between reticulate and divergent evolution (i.e., the tokogeny-phylogeny interface *sensu* Hennig 1999). Species delineated using this approach thus represents unique and genealogically exclusive groups of individuals.

Although genealogical concordance was first applied to animal species (Rising and Avise 1993) it has since also been widely used for the delineation of fungal species. Taylor and co-workers (2000) reviewed several of the initial fungal examples and discussed several of the arguments against this approach. In 1991, Dykhuizen and Green indicated that this approach could also be used for asexual microbes when they applied genealogical concordance to demonstrate the impact of recombination on the genetic structure of a bacterial species like *E. coli*. Although this approach has been widely used in the field of taxonomy it is not often used in bacterial systematics. This is despite the fact that bacterial species are considered to be unique and exclusive groups (Achtman and Wagner 2008) which is in agreement with the principles of the genealogical concordance approach for delineating species.

Practically, genealogical concordance analyses *sensu* Baum and Shaw (1995) involve the use of sequence information for multiple independent loci. These sequences may be obtained by mining the genomes of populations of the organisms forming the subject of the investigation (focal organisms) and relevant outgroups or by making use of the information utilized in typical MLSA studies (i.e., the sequences for 3-7 independent loci) (Gevers et al. 2005). Concordance among the phylogenies of these loci is then investigated so as to delineate groups that are representative of putative species. Here, the aim is to find the point in the individual trees where the genealogies pass from being concordant to discordant (population-level reticulate processes) (Figure 1B). Among the focal organisms, this point in the genealogies delimits the group of individuals that likely form part of the same species.

The genealogical concordance approach differs markedly from MLSA where species delineation is based on the phylogeny of the concatenated sequence dataset. More importantly, however, MLSA does not provide information on the relative position of the species boundary (i.e., some level of subjective interpretation is typically needed to demarcate the potential species boundary in MLSA trees). By contrast, the species boundary is clearly evident from the results of genealogical concordance analyses (compare Figure 1A and B). Different from MLSA, the genealogical concordance approach thus provides an objective way of generating a plausible species hypothesis for the organisms of interest.

A taxonomic workflow that includes genealogical concordance analysis

Most bacteriologist and bacterial systematists are confronted at some point or another with the task of identifying species from among a set of focal organisms. When the species are known to Science, the task is often much simpler and more straightforward than when one is dealing with novel taxa. In such cases, the use of the integrated set of procedures associated with polyphasic taxonomy has shown its value, time and again (Stackebrandt et al. 2002). Given its ability to objectively generate plausible species hypotheses, inclusion of genealogical concordance in the taxonomic workflow would thus improve the correlation between described species and those occurring in nature. The analysis of objectively identified and plausible species hypotheses would also provide a sound theoretical basis to studies that aim to integrate diverse types of biological data to delimit stable taxonomic units (Dayrat 2005; Padial et al. 2010). Below we outline how the principles of genealogical concordance may be integrated with existing taxonomic approaches by providing the six fundamental steps of a typical workflow.

Step 1: Collect the focal organisms

The general aim during this step of the process is to obtain a group of individuals that are representative of the diversity of the focal organism(s). The rationale is that use of a sufficiently representative set of individuals allows for delineation of a species hypothesis that approximates that of the group found in nature. Failure to use a representative sample of individuals during any species identification procedure results in the diagnosis of groups that represent sub-populations of the real species. This type of sampling inadequacy has been referred to as the "iceberg bias" (Tibayrenc 1999) and had been suggested to account for much of the taxonomic instability observed for microbial species (Leslie et al. 2001).

Step 2: Collection of gene sequence information for multiple independent loci

As mentioned before, the loci typically employed in MLSA are well suited for genealogical concordance studies. Ideally, these loci should be selectively neutral (Taylor et al. 2000). In other words, the genes should encode products that are involved in housekeeping functions and that are not subject to balancing selection (Taylor et al. 2000; Gevers et al. 2005). The loci should also be present in all individuals of the focal organism(s). Additionally, the loci used should evolve independently and not be linked tightly, i.e., they should preferably be encoded at diverse locations on the organism's genome (Gevers et al. 2005). Finally, potential

problems associated with the use of non-orthologous sequences for a specific locus, although impossible to avoid completely (see Step 5 below), may be reduced by employing single-copy genes (Gevers et al. 2005).

Step 3: Inference of single-locus genealogies

For this purpose, any of a range of phylogenetic inference methods are typically used. These include methodologies based genetic distance and maximum parsimony, as well as those that can incorporate suitable models of molecular evolution such as maximum likelihood and Baysian inference (Holder and Lewis 2003). Additionally, for all the inferred genealogies, branch support should be estimated by making use of bootstrapping or some other appropriate procedure (Felsenstein 2004).

Step 4: Comparison of genealogies to identify groups potentially representing distinct species

During this step, individual genealogies are investigated for the presence of consistent groups amongst the focal organisms (Sites and Marshall 2004). Should these organisms represent a distinct species, they will group together in the various genealogies, because they are more related to one another than to non-conspecific individuals included in the analyses. However, the relationships among individuals within such an exclusive group could differ dramatically between the different gene trees, because of the dissimilar evolutionary histories of the loci examined (Baum and Shaw 1995; Sites and Marshall 2004). In Figure 1B, for example, the gene genealogies support two such groupings and "Z" indicates the tokogeny-phylogeny interface where their species boundaries are situated. These basal and exclusive groups (i.e., those that do not contain other species and whose members are each other's closest relatives) represent species hypotheses, the probability of which requires evaluation in the final step of this workflow (see below).

When large collections of isolates are included in the study, the comparison of individual gene trees for identifying consistent groups is a daunting task. One way to guide and simplify the process is to generate a consensus tree based on the individual trees (Figure 1B). Some authors suggest the use of a strict consensus for this purpose (Sites and Marshall 2004), while others advocate majority-rule consensus trees (Baum 2007). Another approach is to infer a tree from the concatenated sequences of all the loci examined (e.g. an MLSA tree). From such a consensus tree (typically containing unresolved clusters of individuals) or MLSA tree, clusters of isolates can then be identified, after which their exclusive and basal grouping can more easily be examined manually in the single locus genealogies. In Figure 1A, for example, isolates 1-4 and 5-8 in the

MLSA tree represent two isolate groups whose exclusivity might be evaluated in the gene genealogies presented in Figure 1B.

Although single-gene datasets are notorious for not allowing the inference of robust phylogenies (e.g., Rokas et al. 2003; Gontcharov et al. 2004), the statistical support they do include may in some cases assist in the species delineation process. The premise is that the groups identified (i.e., those potentially presenting species) in the various genealogies should not strongly contradict one another (although relationships within such groups may differ markedly due to population-level processes). For example, depending on the specific dataset, different gene genealogies may support consistent groups without those groups receiving significant bootstrap support; this lack of support would nonetheless not influence the conclusion that all of the loci support the basal and exclusive nature of the delineated group. However, the existence of one or more strongly-supported competing phylogenetic hypotheses in some genealogies, could suggest that the respective loci be examined in more detail. For example, apart from showing that a particular species hypothesis might be invalid, the existence of contradictory clusters of individuals in sets of gene genealogies might also point towards methodological errors or the effects of specific evolutionary phenomena (e.g., De Queiroz et al. 1995; Maddison 1997; Taylor et al. 2000), the effects of which could influence the ultimate species hypothesis (see Step 5 below).

Step 5: Evaluation of potential causes for contradicting groups delineated using genealogical concordance

Groups that are apparently incongruent may be recovered using some of the independent loci utilized for the analysis of genealogical concordance. The generation of the ultimate species hypothesis from such data is, however, still possible with some understanding of the causes of the discrepancies. The four primary sources of discord among loci for identifying species hypotheses (i.e., incomplete lineage sorting, horizontal gene transfer, gene duplication/extinction, non-neutrality) are briefly described below.

Incomplete lineage sorting: This phenomenon, also denoted to as "deep coalescence", refers to the occurrence of the same neutral alleles (in an otherwise polymorphic locus) in distinct species (Maddison 1997). These alleles existed in the ancestral lineage and, after speciation, remained in the extant populations of separate species. The presence of such alleles would cause closely related but non-conspecific individuals to group together. In other words, common ancestry of the affected loci extends deeper than the speciation event, thus causing their evolutionary trajectories to coalesce with that of the ancestral homolog (hence the term "deep coalescence")

(Maddison 1997; Taylor et al. 2000). Compared to all of the loci in an individual's genome, however, those affected by incomplete lineage sorting could be expected to be few (Taylor et al. 2000; Galtier and Daubin 2008).

Horizonal gene transfer (HGT): The evolutionary histories of all bacteria is to some extent influenced by HGT (Ochman et al. 2000). Just as loci acquired from sources outside the species would cause isolates or species to occupy spurious positions in gene trees (Philippe and Douady 2003), their use will also affect the results of genealogical concordance analyses. Although the emphasis on housekeeping loci during the second step of this workflow minimises the chances of accidentally utilizing loci prone to HGT, housekeeping genes can also be impacted by HGT (Boucher et al. 2001). Generally, however, not many housekeeping loci are expected to be characterized by HGT-derived evolutionary trajectories. This is in agreement with the prediction of the well-known "complexity hypothesis" (Jain et al. 1999) that genes encoding products which interact with numerous other gene products (a class of genes to which many housekeeping loci belong) are less likely to experience HGT.

Gene duplication/extinction: Ancestral duplication of a locus, followed by complete or interrupted paralog loss during the divergence of populations could lead to the presence of non-orthologous versions of the locus in extant populations of different species (Maddison 1997). Therefore, the coalescence of non-conspecific individuals' genealogies for these loci will predate the divergence of the species. In other words, the use of such loci for genealogical concordance analyses will have the same effect as incomplete lineage sorting.

Non-neutrality: Although the second step in this workflow endeavours to ensure the use of neutrally evolving loci for use in genealogical concordance analysis, non-neutrality of loci is difficult to completely exclude. This is mainly because the evolutionary processes governing the emergence and maintenance of a species are unique and specific to that particular species. Also, the selection of loci for use in genealogical concordance is usually not based on empirical data for the focal organisms, but rather is guided by what is known from the general literature (hence the avoidance of loci known to be involved in phenotypes that potentially confer selective advantages) (Taylor et al. 2000; Gevers et al. 2005). Nonetheless, strongly non-neutrally evolving loci might inadvertently be included in the analysis for genealogical concordance. In such genealogies, groupings among individuals can be expected to correspond with the selection experienced by the specific locus and not

speciation (Taylor et al. 2000). In some cases, such loci might be identifiable by the inordinately long coalescence times (i.e. long branches) for the relevant groupings (Taylor et al. 2000).

Clearly, all four of the phenomena listed above can impact significantly on how genealogical concordance data are interpreted. However, their occurrence is a normal part of the evolutionary processes of all organisms. Consequently, genealogical concordance data for any taxon or group of taxa may be realistically expected to include signatures of one or more of these phenomena. The apparent incongruence among some loci can thus not be viewed as grounds for rejecting a particular species hypothesis. In fact, Taylor et al. (2000) argued that "*a few loci, or even one, that shows fixation in one or the other of the phylogenetic species is evidence of genetic isolation*". It also follows that large numbers of independent loci need not necessarily be required for the generation of species hypotheses using genealogical concordance analysis (Taylor et al. 2000).

Step 6: Evaluation of corroborative evidence for the species hypotheses

Up until the previous step of the workflow, all of the exclusive and basal groups are purposefully referred to as "species hypotheses". Due to the uncertainties introduced by the occurrence of evolutionary phenomena that might bias conclusions from genealogical concordance data, this final step of the workflow aims to test these hypotheses using additional biological and genetic evidence. These can include some of the criteria traditionally included in bacterial systematics studies such as DNA-DNA hybridization, typing methods and ecological information (Stackebrandt et al. 2002; Rossello-Mora and Amann 2015). In the postgenomic era, however, a whole range of additional properties are increasingly used as independent lines of evidence for supporting species hypotheses. These include various genome similarity criteria such as ANI (Konstantinidis and Tiedje 2005). The information encoded on these genomes can also be interrogated for the occurrence of so-called "synapomorphic" traits (shared derived traits) that characterise all the members of the delineated species. Overall, the use of such diverse criteria for characterising bacterial species is thus fully congruent with the traditional spirit of polyphasic taxonomy were multiple types of genetic and biological data are used (Vandamme and Peeters 2014).

Delineation of *Pantoea allii,* dealing with biologically similar species

When applying the current polyphasic approach it is often difficult to find differences to support the separation of two phenotypically similar species as was the case with *Pantoea allii*. *Pantoea ananatis* is often isolated as

an epiphyte and in some cases as a pathogen of onions (Coutinho and Venter 2009). Various isolates from onion were found to be similar to *P. ananatis* based on phenotype and 16S rRNA sequences, but grouped separately from *P. ananatis* using MLSA (Brady et al. 2008). Later, with the application of polyphasic taxonomy, using DNA-DNA hybridisation, certain phenotypic traits, amplified fragment length polymorphism (AFLP) and MLSA, the bacteria from onion were shown to represent a new coherent species, distinct from *P. ananatis*, and described as *P. allii* (Brady et al. 2011). This study reconfirmed that 16S rRNA sequence data often provide limited resolution and has limited value when delineating species within genera belonging to the *Enterobacteriaceae* (Brady et al. 2008; Brady et al. 2013).

Apart from the uncertainties associated with recovering a species hypothesis from an MLSA tree, its overall clustering pattern may also be influenced by the data used. When sequences for different genes are combined to generate a tree, phylogenetic signal from one or a few of the sequences employed might dominate or mix in such a way so as to produce a tree not resembling any of the possible underlying evolutionary histories (Baum 2007; Salichos and Rokas 2013; Thiergart et al. 2014). We therefore interrogated the separation of *P. ananatis* and *P. allii*, which are closely related species that often occur in sympatry, by making use of genealogical concordance analysis. To accomplish this, we utilised the taxonomy workflow presented above. First, a collection of isolates representing the known diversity of both taxa were assembled (Step 1). The sequences for a set of independent loci were then obtained (Step 2). For this purpose, the original MLSA dataset, consisting of sequences for the genes *gyrB, infB* and *atpD*, were expanded to include those for two additional independent loci, *ompF* (encoding the major outer membrane protein) and *pmrB* (encoding a sensor kinase that forms part of a two-component regulatory system). The genealogies for these loci were inferred with maximum likelihood analyses using appropriate model parameters (Step 3). See Supplementary File 1 for information on isolates, sequencing and accession numbers for *ompF* and *pmrB*.

Comparison of the individual genealogies for the five loci showed that the *P. ananatis* and *P. allii* isolates formed consistent groups (Step 4 of the workflow; Figure 2). No discord was detected among the groups delineated with these loci (i.e., Step 5 of the workflow was not required). These groupings thus suggested that the MLSA tree produced by Brady et al. (2011) likely represented an average of the phylogenetic signal incorporated in all three the loci originally used to generate it (Baum 2007). However, all five of the examined loci supported incongruent within-group relationships (Figure 2), which suggested considerable population-level

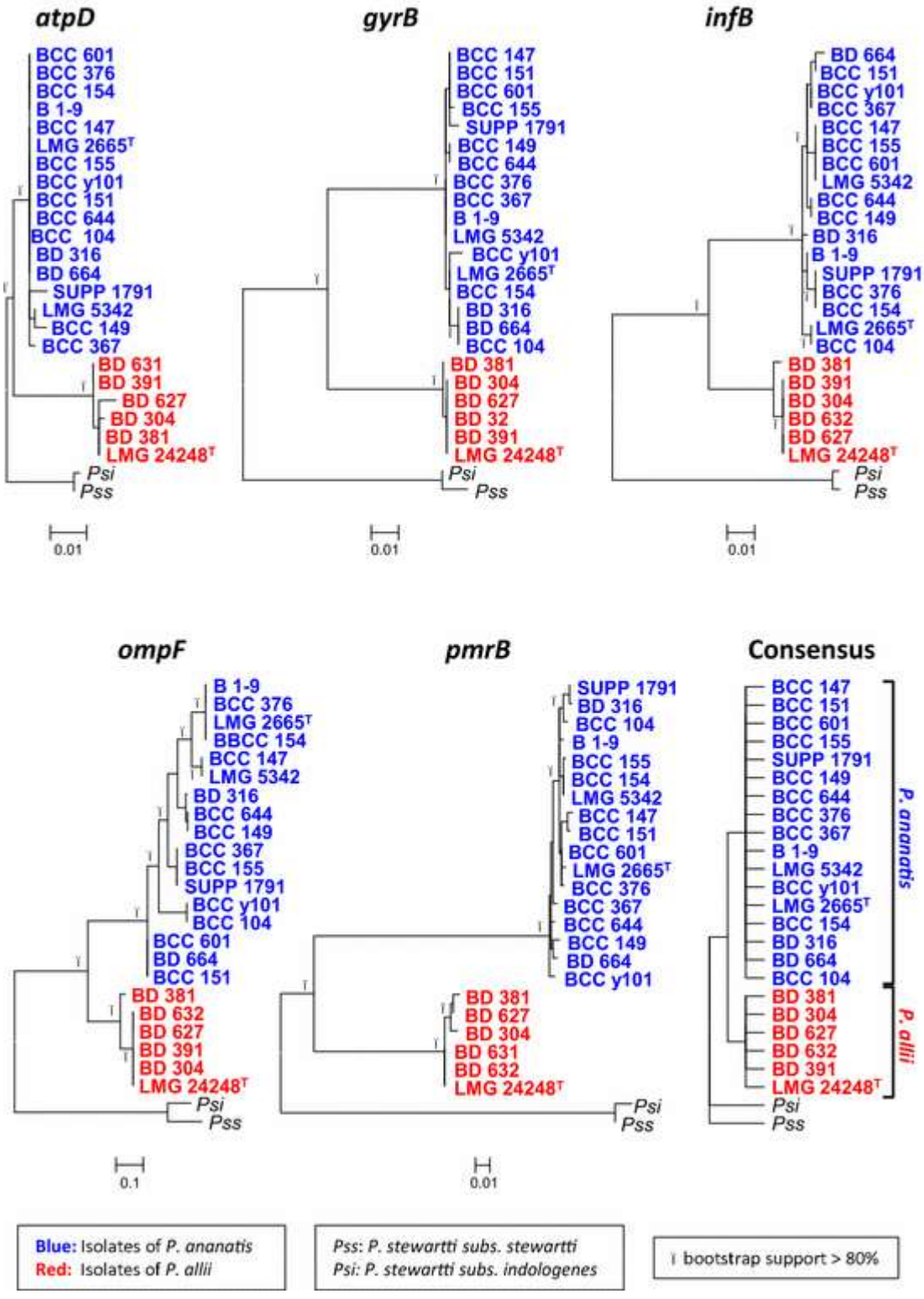


Figure 2. Maximum likelihood phylogenies inferred from the sequence data of the *atpD, gyrB, infB, ompF* and *pmrB* genes for isolates identified as belonging to *Pantoea ananatis* (17 strains) and *P. allii* (6 isolates). *P. stewartii* subs *stewartii* and *P. stewartti* subs. *indologenes* were used for outgroup purposes. Bootstrap values (≥ 80%) are indicated with dots at the respective branches. The consensus gene tree is based on the same set of 5 genes. The scale bars indicate the number of nucleotide changes per site. See Supplementary Dataset F1 for GeneBank accession numbers.

reticulation in both groups. This also caused polytomies (i.e., multifurcating branches) in the strict consensus of the five genealogies (Figure 2). Following the principles of genealogical concordance, these groups thus represent basal and exclusive genealogical species.

The existence of *P. ananatis* and *P. allii* is supported by additional lines of evidence (Step 6 of the workflow). As was reported previously (Mergaert et al. 1993; Brady et al. 2011), both species are supported by DNA-DNA hybridization (i.e., relative binding ratios of 90 - 99 % and 85 – 100% among isolates of *P. allii* and *P. ananatis*, and 55% between the type strains of the two species). The same pattern was also apparent using ANI (i.e. of 99% amongst isolates of *P. ananatis*, and 88% between the *P. ananatis* and *P. allii* type strains) (Supplementary File 2). We also investigated the separation of these two species based on the genome-wide similarity of their shared gene content by making use of a recruitment plot (Ghai et al. 2010) (Figure 3). This summary of the distribution and similarity of shared genes showed that, despite sharing a large number of genes, the similarity of their shared genes seldomly exceeded 90% (Figure 3). Taken together, all of these data confirm the split between *P. allii* and *P. ananatis* and show that both indeed represent valid species.

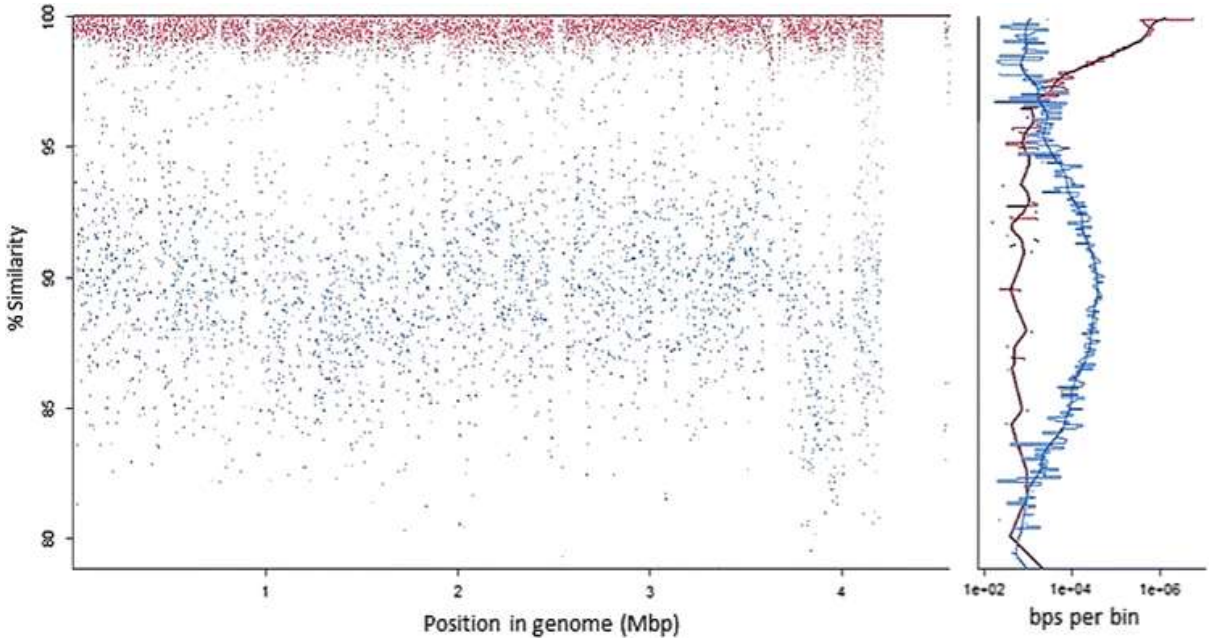


Figure 3. Recruitment plot comparing the genome sequences of *Pantoea ananatis* strains (LMG 20103 and PA13) and the *P. allii* type strain (LMG 24248) with the *P. ananatis* (AJ3355) genome as reference. Genes belonging to the *P. ananatis* strains are indicated in red and *P. allii* genes are indicated in black. Genes are ordered on the x-axis according to their position on the reference genome. The corresponding position of the gene on the y-axis indicates the gene's similarity with the corresponding gene on the reference genome. The side graph indicates the distribution of genes according to their similarity values.

Delineation of *Paraburkholderia kirstenboschensis*, dealing with a diverse species where some taxonomy metrics produce borderline values

During a survey of the rhizobia associated with papilionoid legumes indigenous to the Cape Floristic Region in South Africa, various strains belonging to the genus *Burkholderia sensu lato* were isolated. Several of these isolates formed a unique but diverse cluster based on their 16S rRNA sequences (Beukes et al. 2013). Although these isolates were well separated from the validly described species, their possible status as a single new species was investigated using the principles of genealogical concordance (Steenkamp et al. 2015). This resulted in the description of *B. kirstenboschensis* (Steenkamp et al. 2015), which was recently renamed as *Paraburkholderia kirstenboschensis* (Dobritsa and Samadpour 2016).

The process used for the delineation of *Par. kirstenboschensis* mirrored the taxonomy workflow presented above (Steenkamp et al. 2015). The taxa considered in the study included all of the possible individuals potentially representing this species, as well as isolates of closely related species and outgroup taxa (Step 1 of the workflow). For this collection of individuals, the DNA sequences for four independent loci (16S rRNA, *atpD, recA* and *rpoB*) were obtained (Step 2) from which gene trees were inferred using maximum parsimony (Step 3). Interrogation of the results (Steps 4 and 5), revealed that the *Par. kirstenboschensis* isolates formed a basal and exclusive group in all four of the genealogies, of which the strict consensus tree suggested population-level reticulation (Supplementary File 3). However, this group lacked statistical support in the 16S rRNA tree. Problems associated with the use of 16S rRNA sequences in bacterial systematics are well documented (See Steenkamp et al. 2015) and the lack of support was likely caused by the limited phylogenetic signal included within these sequences. This limited signal is most probably dominated by the ancestral characters, which in the case of *Par. kirstenboschensis*, it shares with its close relatives. Nevertheless, these data allowed for the delineation of a single, strong species hypothesis for *Par. kirstenboschensis*.

Various lines of evidence were investigated to evaluate the hypothesis (Step 6). Among the *Par. kirstenboschensis* isolates, ANI values of >96% were obtained, while comparisons with other species generated values of <92%, which falls well in the range of what is typically obtained from comparisons among conspecific individuals and separate species (Goris et al. 2007; Richter and Rossello-Mora 2009). A similar trend was also observed when gene content and similarity were considered using a recruitment plot. Most of the shared genes between isolates of *Par. kirstenboschensis* were highly similar as opposed to the lower similarities observed for

the genes shared with the closely related *Par. caledonica* (Supplementary File 4). Steenkamp et al. (2015) also used comparative genomics to identify a range of biological processes that potentially represent synapomorphic properties for *Par. kirstenboschensis*. However, in contrast to these different lines of evidence for supporting *Par. kirstenboschensis*, DNA-DNA hybridization did not convincingly support it. Although most of the within-species comparisons produced reassociation values above the 70%, those involving comparisons between isolates with very different genome sizes produced values below the widely recognized species threshold (Steenkamp et al. 2015).

The description of *Par. kirstenboschensis* demonstrated that all of the usual "boxes" do not necessarily need to be "ticked" for a new species to be recognized. After objectively generating a credible "species hypothesis", various lines of evidence were shown to support it. Even if the so-called gold standard for bacterial taxonomy (Rossello-Mora and Amann 2015) did not support the hypothesis, it also did not disprove it. An understanding and logical explanation of what caused the lower DNA-DNA reassociation values informed the integration of these findings with the original hypothesis (see the Steenkamp et al. paper and the discussion therein). Systematic documentation of Earth's bacterial diversity will undoubtedly lead to the discovery of numerous bacteria where application of the usual taxonomic metrics would fail to delineate real and objective biological units. Their ultimate description will thus be dependent on the use of an approach in which the use of arbitrary metrics is de-emphasized.

Delineation of *Escherichia coli* Clades, dealing with recently evolved species

In 2009, Walk et al. described five *Escherichia* clades (I-V) within *E. coli sensu lato* based on an extended multilocus sequence typing (MLST) scheme utilizing the sequence data for 22 independent housekeeping loci. Despite being phylogenetically distinct, these clades could not be separated based on the standard set of biochemical and enzymatic reactions used for the identification of species in the *Enterobacteriaceae*. Subsequent genome-based studies (Luo et al. 2011) suggested that *Escherichia* Clades III, IV and V represent environmentally adapted species, while *Escherichia* Clade I is a member *E. coli sensu stricto* where it likely represents one of the various phylogroups of this taxon (Clermont et al. 2011). In this paper, we investigated the use of genealogical concordance analyses to delineate *Escherichia* Clades III, IV and V by using the workflow presented above.

To accomplish the first three steps, maximum likelihood genealogies were inferred for the 22 housekeeping loci by making use of the published dataset of Walk et al (2009) (see Supplementary File 5). Comparison of these single gene trees (Step 4) revealed that the individuals belonging to *Escherichia* Clade V formed a distinct and consistent group for 19 of the loci examined (Supplementary File 5). Furthermore, *Escherichia* Clade III was recovered as a distinct and consistent group in 14 of the gene trees and *Escherichia* Clades IV from another set of 12 gene trees (Supplementary File 5). The consensus of these genealogies also suggested significant population-level reticulation within each of the groups (Figure 4).

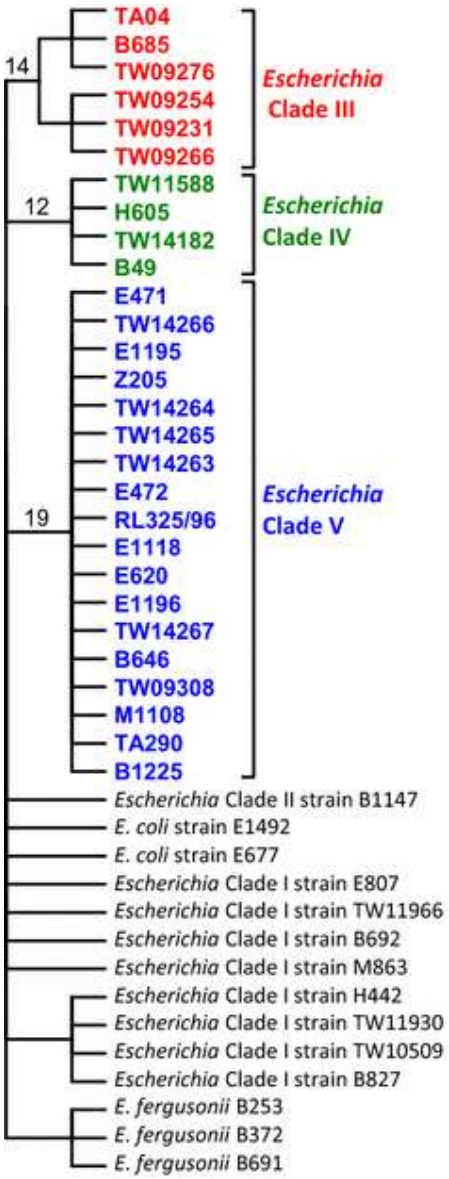


Figure 4. A Majority-rule consensus tree summarising the individual gene trees for *Escherichia coli sensu stricto* and the *Escherichia* Clades described by Walk et al. (2009). The indicated polytomies are indicative of reticulate evolution observed for most of the 22 genes included in the original extended multilocus sequence typing (MLST) analysis. Numbers indicated on branches indicate the number of genes out of the 22 gene genealogies supporting the respective clusters.

Interrogation of the possible causes of the discordant clustering in some of the 22 genealogies (Step 5), indicated that HGT was the likely reason for not recovering a monophyletic *Escherichia* Clade V using the *dnaG*, *torC* and *recA* sequences (Figure 4 and Supplementary File 5). HGT probably also caused the discordant clustering of individuals of *Escherichia* Clade III (e.g., *torC*, *dnaG*, *purA*, *grpE*, *adk* and *aroE*) and *Escherichia* Clade IV (e.g., *torC*, *dnaG* and *aroE*). In all three of these cases, a small number of individuals from the respective Clades formed part of assemblages primarily including individuals representing *E. coli sensu stricto*. This pattern of HGT is not unexpected as genetic exchange among individuals of *Escherichia* Clades III-V and *E. coli sensu stricto* may occur under natural conditions where these bacteria often exist in the same environment (Berthe et al. 2013).

Some of the discordant clustering patterns observed for the individuals of *Escherichia* Clades III and IV is probably also attributable to incomplete lineage sorting (or ancient duplications/extinctions). For example, in some genealogies, coalescence of these Clades seem to predate their divergence (e.g., *kdsA*, *lysP*, *mdh*, *arcA* and *mutS*) or their divergence from *E. coli sensu stricto* and the other clades (e.g., *icdA*, *gyrB* and *grpE*). The occurrence of incomplete lineage sorting (or ancient duplications/extinctions) in the loci of such closely related taxa is, however, not unexpected (Galtier and Daubin 2008).

Overall, application of Steps 1-5 our taxonomy workflow suggested that *Escherichia* Clades III-V indeed represent plausible species hypotheses. These hypotheses were also supported by genome-based properties (Step 6 of the workflow). Comparison of the type strain of *E. coli* and strains of Clade V produced ANI values around 90% (Supplementary File 2). The recruitment plot also illustrated this separation between *E. coli sensu stricto* isolates and Clade V based on the similarity of the shared genes (Supplementary File 4). This was also true for *Escherichia* Clades III and IV, which most like share a recent common origin. They respectively shared ANI values of 91.8% and 92.3% with *E. coli sensu stricto* and an ANI of 96.3% with one another (Supplementary File 2), and a similar trend was observed in terms of genome-wide gene content and similarity (Figure 5).

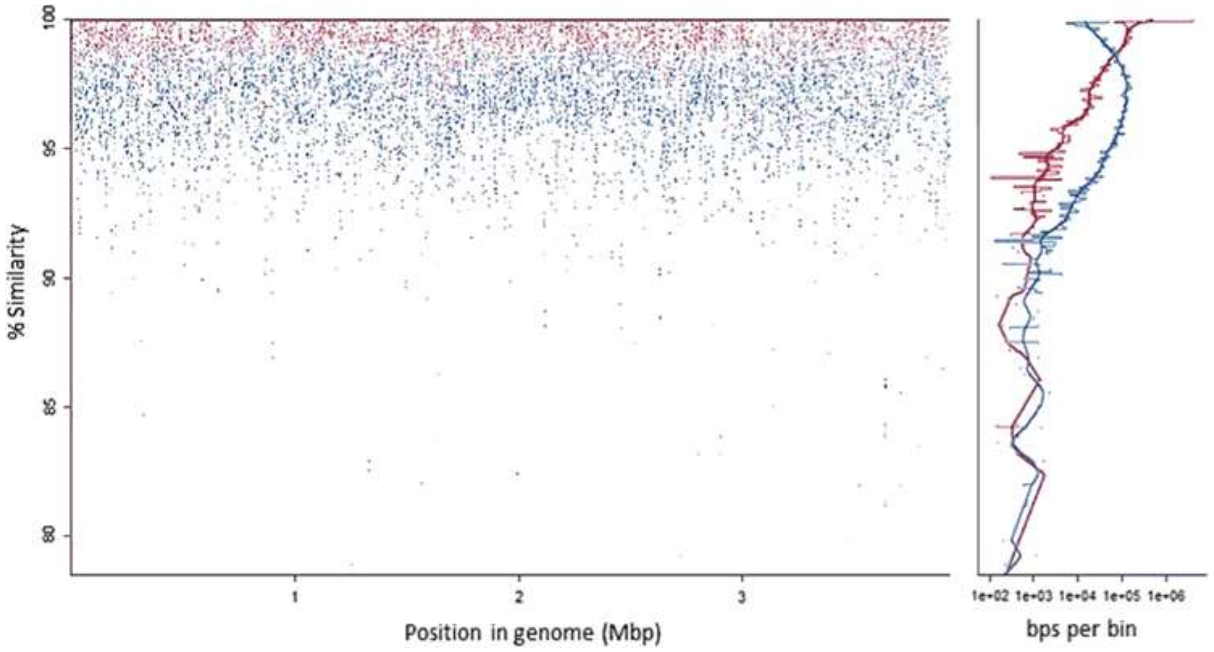


Figure 5. Recruitment plot comparing the genome sequences of an *Escherichia* Clade III strains (TW09231) and two *Escherichia* Clade IV strains (TW14182 and H605) with the *Escherichia* Clade III strains (TW09276) genome as reference. Genes belonging to the *Escherichia* Clade III strain are indicated in red and those belonging to the *Escherichia* Clade IV strains are indicated in black. Genes are ordered on the x-axis according to their position on the reference genome. The corresponding position of the gene on the y-axis indicates the gene's similarity with the corresponding gene on the reference genome. The side graph indicates the distribution of genes according to their similarity values.

Of the three species hypotheses examined here, previous work suggest that *Escherichia* Clade V likely split from the other lineages of *E. coli sensu lato* much earlier than Clades III and IV, which probably diverged much more recently (Wirth et al. 2006). All of the available data suggest that these clades bear the hallmarks of being unique and distinct species of *Escherichia*. The ANI-based similarity of *Escherichia* Clades III and IV is somewhat higher than usually reported for non-conspecific individuals (i.e., 95-96%) (Goris et al. 2007; Richter and Rossello-Mora 2009), but this would be expected for recently evolved nascent species where sufficient time has not yet elapsed for genetic drift to have caused differential fixation at most of their loci (Taylor et al. 2000). For such closely related taxa, clear phenotypic differences may also not be evident, as these typically require substantial evolutionary time to develop, especially when they are not selected for during initial lineage splitting (De Queiroz 2005; Staley 2006).

Detailed analysis of the biological and genetic properties of *Escherichia* Clades III-V will undoubtedly reveal additional independent characters that support their existence as unique species in the genus. Current taxonomic

practice, with its focus on "species coherence" will not easily allow for their formal recognition as species (Brisse et al. 2014; Rossello-Mora and Amann 2015). However, the latter is crucial if we are to improve our understanding of the biological and ecological roles of these bacteria. In other words, the use of species definitions that match or that at least, approximates the units occurring in nature will contribute significantly towards the study of biological systems.

The way forward – revising the current taxonomic approach for the 21st century

Bacterial species are the basic units used by microbiologists when describing the diversity encountered during ecological and clinical studies (Rosselló-Mora and López-López 2008). To be meaningful, a species should be defined by interpreting both its evolutionary history and its biology. Indeed, several calls have been made to adjust microbial taxonomy, specifically the polyphasic approach, in the post-genomic era (Thompson et al. 2013; Vandamme and Peeters 2014) to reflect our increased understanding of bacterial speciation and evolution.

By focusing on the boundary between populations and species, we have shown that genealogical concordance provides an objective evolution-based approach to define potential bacterial species or to produce what we refer to as "species hypotheses". Although the information typically included in taxonomic studies of bacteria can be utilised, our procedure incorporates a method for generating plausible and realistic species hypotheses to be tested further using other lines of evidence. Incorporation of the procedure of delineating putative species into a simple and straightforward workflow streamlines the process of diagnosing biological units that are representative of real bacterial species. This is primarily because the process of diagnosing species is broken down into its logical components (i.e., sample collection, obtaining sequences for multiple independent loci, identifying putative species, and exploring other lines of evidence of corroborating the existence of these species).

The whole concept of using additional lines of evidence for corroborating species hypotheses opens up the field for finding a range of interesting and biologically relevant features unique to a particular species. Apart from the normal genome-based indices typically used, the three examples discussed above demonstrate the value of using an "overview" of overall gene content and gene similarities. This can be done in the form of genome-based recruitment plots that provide visual summaries of both gene content and similarity (Ghai et al. 2010). Additional ecological and, in some cases, geographical distribution data could also be employed to test the

species hypothesis. Also, when phenotypic data is considered it might be informative to include genome-informed phenotypes (Steenkamp et al. 2015) and traits relevant to the biology of the species under investigation. The use of the anthropocentric phenotypic and chemotypic tests (Sutcliffe 2015), which form an integral part of current species description check lists, should be avoided. The exciting challenge to bacterial systematists will be to incorporate combined genome-based comparative and evolutionary approaches in taxonomy in order to study the evolutionary processes and biological constraints underlying the unique and exclusive nature of bacterial species (Baltrus et al. 2016).

Acknowledgements

We are grateful to the South African National Research Foundation (NRF) and the Department of Science and Technology for funding through their Center of Excellence programme.

Dr Seth Walk from the Department of Microbiology and Immunology, Montana State University for sharing the original *E. coli* MLST dataset published in Walk et al. 2009.

References

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431-440

Avise JC, Ball RM (1990) Principles of genealogical concordance in species concepts and biological taxonomy. Oxf Surv Evol Biol 7:23-45

Baltrus DA, Mccann HC, Guttman DS (2017) Evolution, genomics and epidemiology of *Pseudomonas syringae*: Challenges in bacterial molecular plant pathology. Mol Plant Pathol 18:152-168

Baum DA (2007) Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56:417-426

Baum DA, Shaw KL 1995. Genealogical perspectives on the species problem. *In:* Hoch PC , Stephenson AG (eds.) Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St Louis, Missouri, pp 289-303

Berthe T, Ratajczak M, Clermont O, Denamur E, Petit F (2013) Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. Appl Environ Microbiol 79:4684-4693.

Beukes CW, Venter SN, Law IJ, Phalane FL, Steenkamp ET (2013) South African Papilionoid legumes are nodulated by diverse *Burkholderia* with unique nodulation and nitrogen-fixation loci. PLoS ONE 8 (7):e68406

Booth A, Mariscal C, Doolittle WF (2016) The modern synthesis in the light of microbial genomics. Annu Rev Microbiol 70:279-297

Boucher Y, Nesbo CL, Doolittle WF (2001) Microbial genomes: dealing with diversity. Curr Opin Microbiol 4:285-289

Brady C, Cleenwerck I, Venter S, Coutinho T, De Vos P (2013) Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): Proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia* gen. nov as *Lelliottia nimipressuralis* comb. nov and *Lelliottia amnigena* comb. nov., respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter* gen. nov as *Pluralibacter gergoviae* comb. nov and *Pluralibacter pyrinus* comb. nov., respectively, *E. cowanii*, *E. radicincitans*, *E. oryzae* and *E. arachidis* into *Kosakonia* gen. nov as *Kosakonia cowanii* comb. nov., *Kosakonia radicincitans* comb. nov., *Kosakonia oryzae* comb. nov and *Kosakonia arachidis* comb. nov., respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis* nom. nov., *Cronobacter helveticus* comb. nov and *Cronobacter pulveris* comb. nov., respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. Syst Appl Microbiol 36:309-319

Brady C, Cleenwerck I, Venter S, Vancanneyt M, Swings J, Coutinho T (2008) Phylogeny and identification of *Pantoea* species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). Syst Appl Microbiol 31:447-460

Brady CL, Goszczynska T, Venter SN, Cleenwerck I, De Vos P, Gitaitis RD, Coutinho TA (2011) *Pantoea allii* sp. nov., isolated from onion plants and seed. Int J Syst Evol Microbiol 61:932-937

Brisse S, Passet V, Grimont PaD (2014) Description of *Klebsiella quasipneumoniae* sp. nov., isolated from human infections, with two subspecies, *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* subsp. nov. and *Klebsiella quasipneumoniae* subsp. *similipneumoniae* subsp. nov., and demonstration that *Klebsiella singaporensis* is a junior heterotypic synonym of *Klebsiella variicola*. Int J Syst Evol Microbiol 64:3146-3152

Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E (2011) Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. Environ Microbiol 13:2468-2477

Colwell RR (1970) Polyphasic taxonomy of the genus *Vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. J Bacteriol 104:410-433

Coutinho T, Venter S (2009) *Pantoea ananatis*: an unconventional plant pathogen. Mol Plant Pathol 10:325-335

Coyne J, Orr H (2004) Speciation. Sinauer Associates, Sunderland Massachusetts

Dayrat B (2005) Towards integrative taxonomy. Biol J Linn Soc 85:407-415

De Queiroz A, Donoghue MJ, Kim J (1995) Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst 26:657-681

De Queiroz K (2005) Ernst Mayr and the modern concept of species. Proc Natl Acad Sci U S A 102 Suppl 1:6600-6607

Dobritsa AP, Samadpour M (2016) Transfer of eleven species of the genus *Burkholderia* to the genus *Paraburkholderia* and proposal of *Caballeronia* gen. nov. to accommodate twelve species of the genera *Burkholderia* and *Paraburkholderia*. Int J Syst Evol Microbiol 66:2836-2846

Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257-7268

Ereshefsky M (2011) Mystery of mysteries: Darwin and the species problem. Cladistics 27:67-79

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland Massachusetts

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci 363:4023-4029

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, De Peer YV, Vandamme P, Thompson FL, Swings J (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol 3:733-739

Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, Martin J, Verdu M, Deschamps P, Moreira D, Lopez-Garcia P, Mira A, Rodriguez-Valera F (2010) Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J 4:1154-1166

Glaeser SP, Kampfer P (2015) Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol 38:237-245

Gontcharov AA, Marin B, Melkonian M (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the *Zygnematophyceae* (Streptophyta). Mol Biol Evol 21:612-624

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81-91

Hennig W (1999) Phylogenetic systematics, University of Illinois Press, Urbana Illinois

Hey J (2001) The mind of the species problem. Trends Ecol Evol 16:326-329

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275-284

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801-3806

Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567-2572

Leslie JF, Zeller KA, Summerell BA (2001) Icebergs and species in populations of *Fusarium*. Physiol Mol Plant Path 59:107-117

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108:7200-7205

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523-536

Mallet J (2001) Species, concepts of. In: Levin SA (ed) Encyclopedia of biodiversity. Volume 5. Academic Press, New York, pp 427-440

Mergaert J, Verdonck L, Kersters K (1993) Transfer of *Erwinia ananas* (synonym, *Erwinia uredovora*) and *Erwinia stewartii* to the genus *Pantoea* emend. as *Pantoea ananas* (Serrano 1928) comb. nov. and *Pantoea stewartii* (Smith 1898) comb. nov., respectively, and description of *Pantoea stewartii* subsp. *indologenes* subsp. nov. Int J Syst Bacteriol 43:162-173

Naser SM, Dawyndt P, Hoste B, Gevers D, Vandemeulebroecke K, Cleenwerck I, Vancanneyt M, Swings J (2007) Identification of lactobacilli by pheS and rpoA gene sequence analyses. Int J Syst Evol Microbiol 57:2777-2789

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304

Padial JM, Miralles A, De La Riva I, Vences M (2010) The integrative future of taxonomy. Front Zool 7:16

Philippe H, Douady CJ (2003) Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 6:498-505

Retchless AC, Lawrence JG (2007) Temporal fragmentation of speciation in bacteria. Science 317:1093-1096

Richter M, Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126-19131

Rising JD, Avise JC (1993) Application of genealogical-concordance principles to the taxonomy and evolutionary history of the Sharp-Tailed Sparrow (*Ammodramus caudacutus*). The Auk 110:844-856

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804

Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25:39-67

Rossello-Mora R, Amann R (2015) Past and future species definitions for Bacteria and Archaea. Syst Appl Microbiol 38:209-216

Rosselló-Mora R, López-López A (2008) The least common denominator: Species or operational taxonomic units? In: Zengler K (ed) Accessing uncultivated microorganisms. American Society for Microbiology, Washington, DC, pp117-130

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327-331

Sites JW, Marshall JC (2004) Operational Criteria for Delimiting Species. Annu Rev Ecol Evol Syst 35:199-227

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kampfer P, Maiden MC, Nesme X, Rossello-Mora R, Swings J, Truper HG, Vauterin L, Ward AC, Whitman WB (2002) Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043-1047

Staley JT (2006) The bacterial species dilemma and the genomic-phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci 361:1899-1909

Steenkamp ET, Van Zyl E, Beukes CW, Avontuur JR, Chan WY, Palmer M, Mthombeni LS, Phalane FL, Sereme TK, Venter SN (2015) *Burkholderia kirstenboschensis* sp. nov. nodulates papilionoid legumes indigenous to South Africa. Syst Appl Microbiol 38:545-554

Sutcliffe IC (2015) Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again. Front Genet 6:218

Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC (2000) Phylogenetic species recognition and species concepts in fungi. Fungal Genet Biol 31:21-32

Thiergart T, Landan G, Martin WF (2014) Concatenated alignments and the case of the disappearing tree. BMC Evol Biol 14:266

Thompson CC, Chimetto L, Edwards RA, Swings J, Stackebrandt E, Thompson FL (2013) Microbial genomic taxonomy. BMC Genomics 14:913

Tibayrenc M (1999) Toward an integrated genetic epidemiology of parasitic protozoa and other pathogens. Annu Rev Genet 33:449-477

Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kampfer P (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249-266

Vandamme P, Peeters C (2014) Time to revisit polyphasic taxonomy. Antonie van Leeuwenhoek 106:57-65

Vandamme P, Pot B, Gillis M, De Vos P, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407-438

Vandamme P, Moore ER, Cnockaert M, De Brandt E, Svensson-Stadler L, Houf K, Spilker T, Lipuma JJ (2013) *Achromobacter animicus* sp. nov., *Achromobacter mucicolens* sp. nov., *Achromobacter pulmonis* sp. nov. and *Achromobacter spiritinus* sp. nov., from human clinical samples. Syst Appl Microbiol 36:1-10

Vanlaere E, Baldwin A, Gevers D, Henry D, De Brandt E, Lipuma JJ, Mahenthiralingam E, Speert DP, Dowson C, Vandamme P (2009) Taxon K, a complex within the *Burkholderia cepacia* complex, comprises at least two novel species, *Burkholderia contaminans* sp. nov. and Burkholderia lata sp. nov. Int J Syst Evol Microbiol 59:102-111

Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS (2009) Cryptic Lineages of the Genus *Escherichia*. Appl Environ Microbiol 75:6534-6544

Wayne LG, Brenner DJ, Colwell RR, Grimont PaD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Truper HG (1987) Report of the *Ad Hoc* Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Evol Microbiol 37:463-464

Wilson EO, Brown WL (1953) The Subspecies Concept and Its Taxonomic Application. Syst Zool 2:97-111

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60:1136-1151