ORIGINAL ARTICLE

The bacterial pangenome as a new tool for analysing pathogenic bacteria

L. Rouli, V. Merhej, P.-E. Fournier and D. Raoult

Aix Marseille Université, URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, 13005 Marseille, France

Abstract

The bacterial pangenome was introduced in 2005 and, in recent years, has been the subject of many studies. Thanks to progress in next-generation sequencing methods, the pangenome can be divided into two parts, the core (common to the studied strains) and the accessory genome, offering a large panel of uses. In this review, we have presented the analysis methods, the pangenome composition and its application as a study of lifestyle. We have also shown that the pangenome may be used as a new tool for redefining the pathogenic species. We applied this to the *Escherichia coli* and *Shigella* species, which have been a subject of controversy regarding their taxonomic and pathogenic position.

Keywords: Bacteria, bioinformatics tools, comparative genomics, pangenome, pathogenic species
Original Submission: 11 May 2015; **Accepted:** 16 June 2015
Available online 26 June 2015

Corresponding author: D. Raoult, URMITE, UMR CNRS 7278, IRD 198, INSERM U1095, Faculté de Médecine, 27 Bd Jean Moulin, 13385 Marseille cedex 5, France
E-mail: didier.raoult@gmail.com

Introduction

The emergence and development of next-generation sequencing technologies (NGS) made the reconstruction of genomes much easier and more accessible than previously [1]. Concerning the study of bacteria, possession and study of more than ten different genomes from the same species is easy, which provides enough data to perform comparisons [1]. Studies of pangenomes arose from these new possibilities and reflect the notion of bacterial species more accurately [2,3]. It is strongly recommended to include a number of genomes in studies to better identify the diversity and composition of the global gene repertoire [1]. The name was quoted in 2005 by Tettelin *et al.* [4], where a clear definition of the pangenome is given. The pangenome (or supragenome) [5,6] has been defined as the whole gene repertoire of a study group [1,2,7]. In this review,

DEFINITIONS

Term	Meaning
Accessory genome	Not unique but not in the core genome.
Allopatric	Here, means living alone in its ecological niche.
Bad bugs	Most dangerous pandemic bacteria for humans.
Closed pangenome	Finished pangenome in which there is no change when new genomes are added.
COG	Cluster of orthologous groups.
Core genome	The pool of genes common to all the studied genomes of a given species.
CRISPRs	Clustered regularly interspaced short palindromic repeats.
KEGG	Kyoto encyclopaedia of genes and genomes
MLST	Multilocus sequence typing, which is used for the typing of multiple loci in molecular biology. It is based on individual phylogenetic analysis or concatenation analysis of multiple housekeeping genes.
Mobilome	All mobile genetic elements of a genome.
MST	Multispacer sequence typing; based on highly polymorphic non-coding sequences.
NGS	Next-generation sequencing.
Non-virulence genes	Genes associated with non-virulence the deletion of which favours virulence.
Open pangenome	A pangenome increasing when a new genome is added to the pangenome.
ORF	Open reading frame.
Pangenome	The repertoire of genes for a group of genomes.
Panmetabolism	The repertoire of metabolic reactions for a group of genomes.
Panregulon	The groups of genes co-regulated observed by transcriptomics analysis.
Resistome	Set of all encoding resistance genes to other bacteria.
SNP	Single nucleotide polymorphism. Variation of only one base.
Species	A homogeneous group of isolates characterized by a phenotypic and genetic resemblance.
Sympatric	Here, means living in a large community in its niche.
TA modules	Toxin/antitoxin modules.

http://dx.doi.org/10.1016/j.nmni.2015.06.005

we study the notion of the bacterial pangenome, which is rapidly growing today (Box 1).

A pangenome can be defined as open or closed (infinite or finite [9]), according to the species' capacity to acquire exogenous DNA [2,10], to have the machinery to use it [10] and to possess a large amount of rRNA [10]. The open or closed nature of a pangenome is bound to the lifestyle of the studied bacterial species [2,7,10]. Moreover, the allopatric species that live isolated in a narrow niche usually have a small genome and a closed pangenome, because they are specialized [7,10]. Sympatric species, living in a community, tend to have large genomes and an open pangenome, a high horizontal rate of genes transfer and several ribosomal operons [7,10].

These studies pose the question of the nature of bacterial species definition. In contrast to the world of Eukaryotes, where this term has been defined relative to fertility [11], the case of Prokaryotes seems to be much more difficult [12]. Usually, bacterial species are defined on the basis of gene contents, phenotypic characters, the nature of the ecological niche and the 16S ribosomal RNA sequences [11,12]. A bacterial species has been defined as 'a group of isolates which are characterized by a certain degree of phenotypic resemblance, by a level of 70% DNA-DNA hybridization and by an identity of at least 97% between 16S rRNA sequences' [13,14] or, more recently, 98.7% [3]. This definition can be applied globally to obligatory pathogens that live in a very narrow ecological niche [11] (allopatry) [13]; there is no real reason for the different adaptation and diversification processes to result in rather coherent groups at the phenotypic and genetic level so they can be designed as a species. Some authors have defined species based on genomic coherence [13], isolate proximity [12] and the ecological niche [11]. We believe that the pangenome represents a new approach to species definition. Indeed, pangenomic studies offer a rather wide panel of possibilities, like predicting the allopatric or sympatric nature of a bacterium, and precisely determining the genomic contents of a group. Based on such results, it is not unrealistic to consider narrowed and closed pangenomes being defined as a species.

Moreover, as quoted by Dagan and Martin [15], a tree based on only one gene or on whole ribosomal protein-encoding genes is too simplistic and not representative of reality. In

BOX 1. Overview of pangenome chronology

Between September 2005 and March 2013, the pangenomes of 41 bacterial species and 12 bacterial genera were published. Among these species and these genera, Proteobacteria are the best represented with, respectively, 49% and 42%. Among these Proteobacteria, the Gamma-Proteobacteria are over-represented with 75% prevalence. Most of the bacteria studied were pathogenic, such as *Haemophilus influenzae* [5] and *Coxiella burnetii* [8], and/or bacteria of general interest like *Escherichia coli* [16]. There is a wide difference between the number of reported species or genera. Some species or genera were heavily studied, such as *E. coli* [16] and *Staphylococcus aureus* [80]. Tables 1 and 2 summarize the publications. There has been a strong increase in the studies since 2009.

TABLE 1. Summary of all the pangenomes studies about bacterial species

Species	References	Phylum	Class
Escherichia coli	[8,16,65]	Proteobacteria	Gammaproteobacteria
Streptococcus pneumoniae	[8,79]	Firmicutes	Bacilli
Salmonella enterica	[8,36]	Proteobacteria	Gammaproteobacteria
Staphylococcus aureus	[8,80]	Firmicutes	Bacilli
Helicobacter pylori	[8,81]	Proteobacteria	Epsilonproteobacteria
Vibrio cholerae	[82]	Proteobacteria	Gammaproteobacteria
Mycobacterium tuberculosis	[83]	Actinobacteria	Actinobacteria
Yersinia pestis	[8,73]	Proteobacteria	Gammaproteobacteria
Acinetobacter baumannii	[8,84]	Proteobacteria	Gammaproteobacteria
Chlamydia trachomatis	[34]	Chlamydiae	Chlamydiia
Bacillus cereus	[1,8]	Firmicutes	Bacilli
Streptococcus pyogenes	[8,54]	Firmicutes	Bacilli
Listeria monocytogenes	[85,86]	Firmicutes	Bacilli
Haemophilus influenzae	[5]	Proteobacteria	Gammaproteobacteria
Pseudomonas aeruginosa	[87]	Proteobacteria	Gammaproteobacteria
Enterococcus faecium	[88]	Firmicutes	Bacilli
Clostridium difficile	[89]	Firmicutes	Clostridia
Francisella tularensis	[8]	Proteobacteria	Gammaproteobacteria
Campylobacter jejuni	[8,90]	Proteobacteria	Epsilonproteobacteria
Bacillus anthracis	[4]	Firmicutes	Bacilli
Clostridium botulinum	[8]	Firmicutes	Clostridia
Buchnera aphidicola	[8,91]	Proteobacteria	Gammaproteobacteria
Actinobacillus pleuropneumoniae	[92]	Proteobacteria	Gammaproteobacteria
Legionella pneumophila	[35]	Proteobacteria	Gammaproteobacteria
Streptococcus agalactiae	[4,93]	Firmicutes	Bacilli
Streptococcus suis	[94]	Firmicutes	Bacilli
Sinorhizobium meliloti	[66]	Proteobacteria	Alphaproteobacteria
Aggregatibacter actinomycetemcomitans	[95]	Proteobacteria	Gammaproteobacteria
Bifidobacterium animalis	[96]	Actinobacteria	Actinobacteria
Prochlorococcus marinus	[8]	Cyanobacteria	Prochlorales
Ralstonia solanacearum	[97]	Proteobacteria	Betaproteobacteria
Rhodopseudomonas palustris	[8]	Proteobacteria	Alphaproteobacteria
Coxiella burnetii	[8]	Proteobacteria	Gammaproteobacteria
Erwinia amylovora	[98]	Proteobacteria	Gammaproteobacteria
Corynebacterium pseudotuberculosis	[99]	Actinobacteria	Actinobacteria
Lactobacillus casei	[100]	Firmicutes	Bacilli
Salmonella paratyphi	[101]	Proteobacteria	Gammaproteobacteria
Oenococcus oeni	[102]	Firmicutes	Bacilli
Staphylococcus epidermidis	[103]	Firmicutes	Bacilli
Corynebacterium diphtheriae	[104]	Actinobacteria	Actinobacteria
Tropheryma whipplei		Actinobacteria	Actinobacteria

TABLE 2. Summary of all the pangenomes studies about bacterial genus

Genus	References	Phylum	Class
Streptococcus	[93]	Firmicutes	Bacilli
Salmonella	[36]	Proteobacteria	Gammaproteobacteria
Vibrio	[82]	Proteobacteria	Gammaproteobacteria
Pseudomonas	[105]	Proteobacteria	Gammaproteobacteria
Burkholderia	[106,107]	Proteobacteria	Betaproteobacteria
Bifidobacterium	[108]	Actinobacteria	Actinobacteria
Chlamydiae	[34]	Chlamydiae	Chlamydiia
Campylobacter	[9]	Proteobacteria	Epsilonproteobacteria
Listeria	[48]	Firmicutes	Bacilli
Dehalococcoides	[109]	Chloroflexi	Dehalococcoidetes
Mycoplasma	[110]	Tenericutes	Mollicutes
Caldicellulosiruptor	[111]	Firmicutes	Clostridia

contrast, pangenome study with different tools may help to define species. Quantum physics is a rift from classic physics and is known to be unintuitive. In quantum physics, we observe that there is no progressive state for an electron between two orbits, because it performs quantitative leaps. It is also shown that the atom does not act as a classic system, which can exchange energy continually.

These physical phenomena fit our definition of the species description used here. Indeed, when we studied the pangenome and we calculated the core/pangenome ratio on theoretically identical species genomes, we did not always obtain a linear graph as expected, instead we saw a break event. When the break is clear, we may conclude that we are faced with two different species.

Here we will present the various methods of analysis, the bioinformatics and experimental tools and the link between pangenome, lifestyle and taxonomy.

Tools

Choice of study subjects

Number of species. We selected 27 bacterial species and compared the core/pangenome ratio depending on the number of tested genomes (Fig. 1) to find the minimum number of genomes necessary for a comprehensive analysis. We noted that in the case of a very closed pangenome (core/pangenome ratio between 100% and 98%), two genomes may be sufficient, and for a closed pangenome six strains seemed sufficient. For an open pangenome, it is more difficult to determine this number of necessary strains. If the pangenome is large, precise analysis can be possible on the basis of ten strains, but in the case of an infinite open pangenome, it is not possible by definition to close it (Fig. 2). This questions the reality of a species such as *Escherichia coli*, for example.

Which strains?. Once the number of isolates has been defined, it is necessary to carefully select strains. Several criteria may be considered. First, if the study involves a pathogen, it can be relevant to include the clinical aspects, as different strains of the same species can cause different diseases. This is the case for *E. coli* [16], where commensal and pathogen isolates can be selected. Among the pathogenic strains, five different clinical groups [16] were selected. Strains also frequently have different geographical origins, like *Coxiella burnetii* and *Yersinia pestis*. These 'geotypes' are usually related to genotypes. A genotype can be defined by different methods: pulsed field gel electrophoresis [17], multilocus sequence typing [18], multispacer sequence typing [19–21] or single nucleotide polymorphisms (SNPs) of the core genome [22]. For *C. burnetii*, every multispacer sequence type is defined by 10 'cox' sequences. Finally, it is also possible to use the phenotype including antibiotic resistance or in stress conditions. These four criteria open a wide range of possibilities and it is interesting to select a large panel of strains to describe the pangenome diversity.

Interest of new species analysis. The real-time genomic base been used during epidemics to discover why and how an isolate was able to cause such an event and, at best, to be able to identify

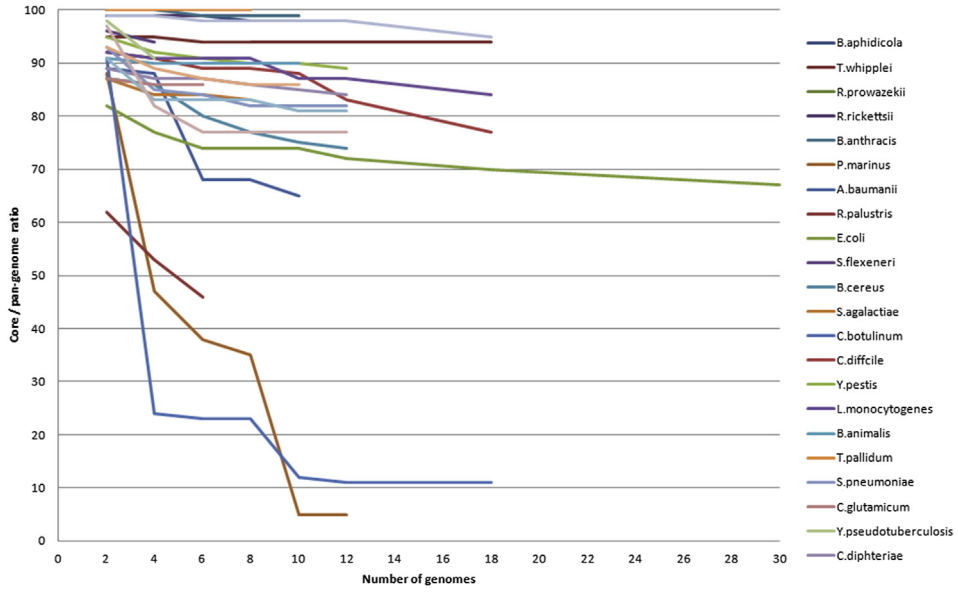

FIG. 1. Study of the core/pangenome ratio function of the number of genomes added in several bacterial species. A closed pangenome is defined when reaching a plateau.

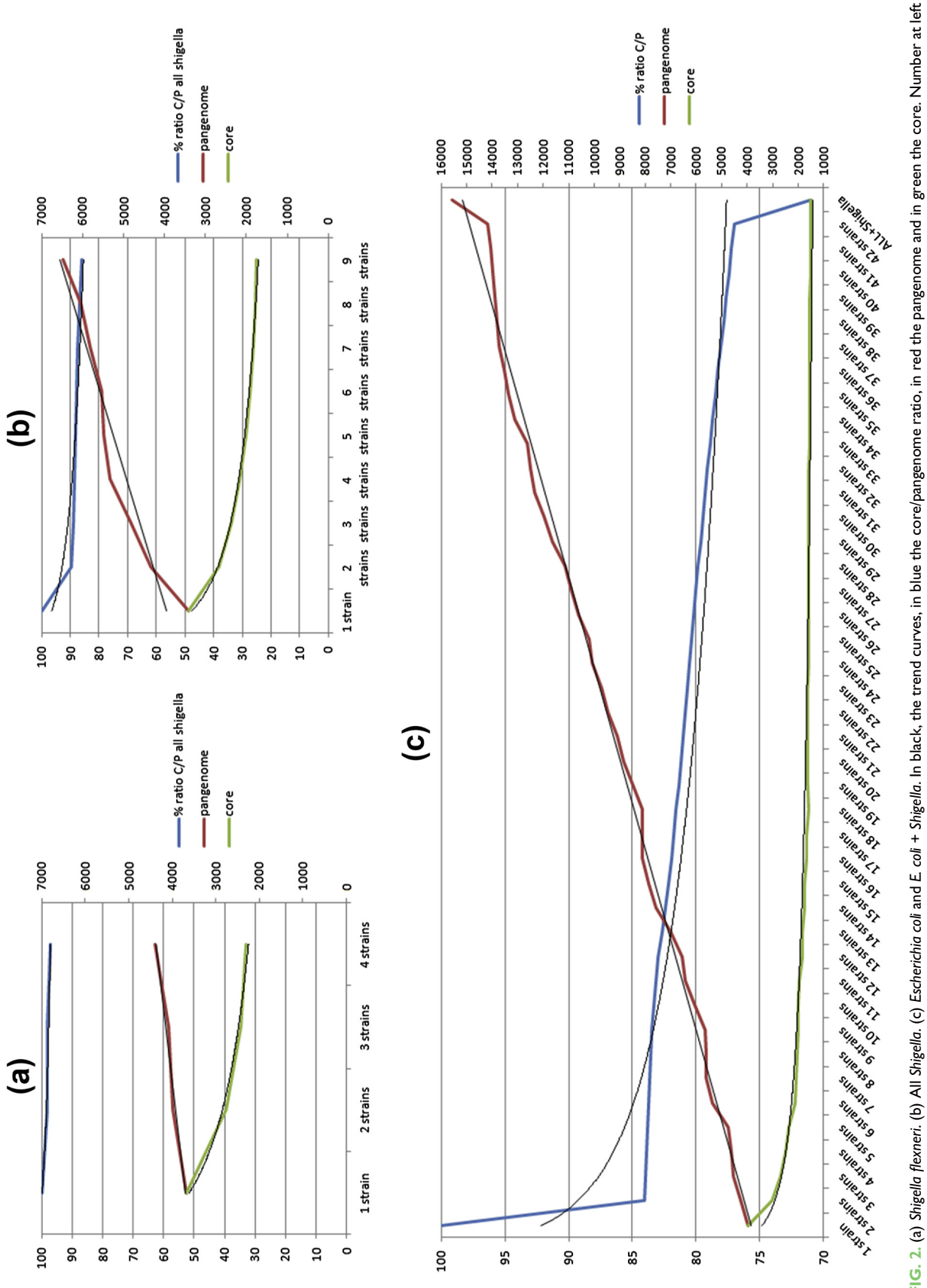

FIG. 2. (a) *Shigella flexneri*. (b) All *Shigella*. (c) *Escherichia coli* and *E. coli* + *Shigella*. In black, the trend curves, in blue the core/pangenome ratio, in red the pangenome and in green the core. Number at left corresponded to percentage and number at right corresponded to number of genes.

specific genetic markers. There are two recent examples of public health use of pangenome analysis: the pandemic in Haiti caused by *Vibrio cholerae* [23] and the German epidemic caused by *E. coli* [24]. Respectively, 23 and 40 genomes were used for these analyses of comparative genomics. During these studies, authors determined the gene content and they placed the isolates of interest in a biotype [23] or an existent pathotype [24]. In the case of *V. cholerae*, all the Haitian clones were clearly related to Nepal [23]. The *E. coli* isolate from the German epidemic was an emerging clone clustering with an enteroaggregative *E. coli* pathotype [24].

Microarrays

Chip technology entered during the pangenomic era, and new tools for designing probes were created. In 2007, PRODESIGN [25] was put into circulation. It is a free online tool (http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html) that can be used to select probes in order to detect the members of gene families in environmental samples. This allows the detection of several gene families simultaneously and specifically in one or several genomes. Moreover, the length and temperature of the probes does not need to be predefined. This tool was, for example, used in a study in 2011 on *Dehalococcoides* [26], to detect and characterize these bacteria in the contaminated sites. A second tool was created in 2009: the PANARRAY [27]. It is a probe selection algorithm that can target several complete genomes with a minimum number of probes. Although microarrays are built on the basis of gene family clusters, PANARRAY uses an approach based and centreed on probes independently of annotation, gene clustering and multialignments. This tool works as well for the known isolates as the unknown; it has been tested on 20 isolates of *Listeria monocytogenes* and also on *C. burnetii* [28,29]. Finally, obtaining data from the microarray approach requires particular and specific analyses, new genes cannot be found. For this purpose, PANCGH [30] was developed in particular in 2009 as well as an associated Web application, PANCGHWEB [31], in 2010. The use of microarrays is only valid for closed pangenomes.

Bioinformatics tools

Composition and annotation. In the first place, searching for orthologues is a crucial step because it allows an estimation of pangenome composition (number of core and secondary genes). To find them, the most commonly used methods consist of one or several sorts of BLAST [32] or OrthoMCL [33]. There are many possibilities available for the annotation step [9,34–36], although COG (Cluster of Orthologous Groups) [37], InterPro [38] and KEGG (Kyoto Encyclopaedia of Genes and Genomes) [39] are the most frequently used. These tools, in particular COG and KEGG, allow a more detailed study of the functional distribution within the core and within the accessory genome. It is possible to look at the difference of distribution in the COG categories [34,35] and at the metabolic pathways [34,35].

Study of the metabolic pathways is not sufficient, however. It is also important and informative to examine protein expression regulation and transcription factors. Moreover, their absence or their presence in one or several isolates can help to explain some isolate characteristics. The online tool P2RP (Predicted Prokaryotic Regulatory Proteins) [40], which became available in 2013, was specially developed to offer a method for simply, quickly and effectively searching for these kinds of proteins that is accessible to all and not only to bioinformatics specialists. The tool covers complete genomes as well as protein sequences and gives detailed and clear outputs.

Alignment and phylogeny. Turning our interest to genome alignment. We can choose a global alignment with MAUVE [41] (or use it for comparison [35]), or we can try a multiple alignment (using CLUSTAL [36]) to perform phylogeny. Most of the time, for phylogeny, MEGA [42] or MAFFT [43] are recommended for tree reconstruction. We can use different algorithms: neighbour joining [36] or maximum parsimony. The search for SNPs in the core genome can be used to estimate the age of species of interest [44]. However, for this kind of analysis it is necessary to possess genomes of very close species to be able to produce a phylogenetic tree and study in detail the mutational events that led to the separation into two different species. This kind of work has been carried out on *Y. pestis*, in which a comparative analysis was conducted with *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* [45].

Resistome and mobilome. To study the resistome, there are databases such as the ARDB (Antibiotic Resistance Genes DataBase) [46], which can be used to look specifically for genes of resistance present in isolates of interest. This database was used, for instance, for *Mycobacterium tuberculosis* [47].

Finally, it is also important to study the mobilome [48]. This represents the set of all the mobile elements (and hence selfish genes) contained in the studied genomes. Generally, we look for the clustered regularly interspaced short palindromic repeats (CRISPRs) with CRISPRs finder [49], phages with PHAST [50] or RAST [51] and insertion sequences with IS finder [52].

Dedicated software. The increase in the number of pangenomic studies led to the development of automated tools, which are more or less specialized. The first one, PANOCT [53] (pangenome orthologue clustering tool), is a tool completely dedicated to orthologue searches. There is no online version, but the source code is available at http://panoct.sourceforge.net/. *Acinetobacter baumannii* isolates were tested and compared with

other tools for orthologue detection. For paralogue detection, PanOCT comes first in terms of accuracy and absence of errors [53]. A second tool, less specialized, the PGAP (pangenomes analysis pipeline) [54], offers the user the possibility to obtain five types of data: clusters of gene functions, species evolution, pangenome profile, and the genetic variation of functional genes. This automatic pipeline, tested on *Streptococcus pyogenes*, is interesting because all the analyses are performed through a single line of command; moreover, it is possible to adapt the parameters to one species. Finally there is the Panseq tool (pangenome sequence analysis program) [55], an online tool (http://lfz.corefacility.ca/panseq/) that allows the user to proceed with three sorts of analyses: search for new regions, allowing the detection of unique zones; analysis of the core and the pangenome, giving information about the SNPs in the core or the distribution of accessory genes; and, finally, a selector of loci allowing us to find those discriminating between selected genomes.

Pangenome Composition

A pangenome is usually divided into three parts [1,2,7]: the core genome, gathering all genes common to all strains of the study; the secondary, called the accessory genome [1,2], which contains genes present between two and $n-1$ strains; and the unique genes, which are present only in a single strain. Inside the pangenome, we can study different features such as resistome, the mobilome and the global metabolism.

Toxin/antitoxin systems

Toxin/antitoxin (TA) genes are small genetic elements that are divided into five groups [56], based on antitoxin nature (small RNA or small protein) and on the interaction type. The type II TA module is the most studied. TA-toxins target different cellular processes depending on their type: ATP synthesis, translation, replication (type II), cytoskeleton (type IV) and peptidoglycan synthesis (type II) [56]. TA modules have different functions, for instance plasmid stabilization and, in the chromosome, mediation of superintegron stabilization [56]. Superintegrons often encode proteins with an adaptive function like virulence, resistance and often contain TA modules. They are also toxic for the host of the bacteria [57]. Comparison of the 'bad bugs' against control species showed that pathogenic capacity is not due to 'virulence factors' (which are periodically, very often, more numerous in non-pathogenic bacteria [58]), but due to a virulent gene repertoire caused by a reduced genome repertoire [59]. 'Virulence factors' is a misleading definition, except for toxins, which may have a direct effect [59]. In 2011, for the first time [60], TA modules were correlated to the pathogenics of some bacteria. Indeed, most of the bad bugs contained significantly more TA modules than their controls [60].

Non-virulence genes

Non-virulence genes are part of an emerging concept where gene expression decreases virulence in the ancestor, and they are lost in pathogenic strains [61]. Their deletion is associated with increased virulence. Originally identified in *Shigella* [62] (lysine decarboxylase), a non-virulence gene may help explain pathogen evolution. It has been described later [62] in *Salmonella*, *Y. pestis* and *Francisella tularensis*. Non-virulence genes can have different roles and be involved, for instance, in metabolism and biofilm synthesis [62]. There are 12 well-known non-virulence genes. A detailed definition of what a non-virulence gene is and what it is not has been proposed [62]. Globally, suppressors and non-functional genes in the ancestor are not, whereas deleted, inactivated or differentially regulated genes may be candidate non-virulence genes. To identify putative non-virulence genes, a reference genome is needed. Then, a very detailed genomic analysis is required on all the sequenced strains [62].

Resistome

Resistome is the term used to indicate all the resistance mechanisms that can be found in an organism [47,63,64]. In a recent study [64], the resistome of 412 multi-resistant bacteria found in four cultivable grounds, four urban soils and two pristine environments was performed, testing 23 antibiotics, considering the large amount of resistant pathogenic isolates [63]. This kind of study was carried out for *M. tuberculosis* in 2013 [47]. The emergence of multidrug-resistant strains prompted the study and 53 genes of resistance have been found, most of these genes (60%) coding for acetyltransferases, having a common ancestral core.

Core and panmetabolism

By analogy with the definitions of the core and the pangenome, the panmetabolism includes all the metabolic reactions that are present in the group of studied organisms, whereas the core contains only the reactions common to all isolates. A complete study was performed on the core and the panmetabolism of *E. coli* [65], including 29 species. The authors found a panmetabolism comprising 1545 reactions, including 885 that belong to the core. The authors noticed that the proportion of core genes and the nature of the pangenome (open or closed) did not reflect panmetabolism distribution. For *E. coli*, for example, known to have an 'infinite' pangenome, they found a large number of core reactions but, as expected, a low number of core genes. They concluded that diversity was lower at gene level than at metabolic level.

Panregulon

Another developed analogy to the pangenome was the panregulon [66]. Studies were either centred more on the core regulon [67] or on the complete panregulon [66]. The panregulon includes all genes controlled by a particular factor of transcription in the studied genomes [66]. In the first work [67], eight isolates of *Listeria monocytogenes* were tested, the core regulon consisted of 63 genes, with a panregulon of 425 genes. In a second study [66] on *Sinorhizobium meliloti* they studied the pangenome and the panregulon at the same time. Based on three isolates, they described a core genome that consisted of 5124 genes and a pangenome of 7824 genes. The panregulon is extremely small compared with the pangenome.

Example of pangenome study: *Legionella pneumophila*

In 2010, using 454 technologies, five complete genomes of *L. pneumophila* [35] were sequenced. It is an intracellular bacterium, a human pathogen that lives in sympatry with other microorganisms within amoebae [68]. *Legionella pneumophila* has an open pangenome. Based on the study of orthologues and helped by BLAST, the core was determined as well as the accessory genome. This was used to describe a core genome that would include 1979 genes, representing 66.9% of the total genome, and a dispensable genome consisting of 978 genes (33.1% of the genome), for which COG categories were assigned. The genome annotation revealed an important number of hypothetical proteins. Most of the genes in the accessory genome belonged to genomic islands, divided into six categories: three different islands connected with drug resistance, one with secretion and transport of heavy metals, three islands with DNA transfer, two CRISPRs systems, seven phage-related systems and 13 islands with no identified function. With regard to these results, authors were able to conclude that the persistence and virulence of *L. pneumophila* is coded by the core genome.

Pangenome for Taxonomy of Pathogenic Species: the Case of *Escherichia* and *Shigella*

Historical taxonomy

For historical reasons related to pathogenicity and particular morphological and biochemical characters, *Shigella* species were classified in a separate genus from *E. coli*. Whereas *E. coli* are usually prototrophic, mobile and ferment many carbohydrates with gas production, *Shigella* are auxotrophic and can produce gas during glucose fermentation. Hence, *Shigella* spp. have many 'negative' characteristics compared with *E. coli*. They are not motile, never grow on the synthetic medium Simmons citrate, lack the activities of phenylalanine deaminase or tryptophan, urease, or lysine decarboxylase, and do not produce H_2S. The division of *Shigella* into four species was based on biochemical and antigenic characterization. These species are divided into serotypes based on a characteristic factor O. However, the distinction between *Shigella* and *E. coli*, especially the enterohaemorrhagic invasive *E. coli* EIEC, is somewhat specious. The O antigens of certain serotypes of *Shigella* are identical or highly related to those of *E. coli*. Like EIEC, *Shigella* causes the dysentery syndrome that consists of fever, diarrhoea with blood, pus and mucus in faeces. The mechanism of *Shigella* pathogenicity is identical to that of EIEC. They enter into epithelial cells to the lamina propria, triggering a major local inflammatory reaction that can lead to abscess formation and ulceration in the colon. *Shigella* should be included in the *E. coli* group. Their individualization was maintained only for practical reasons of medical diagnosis.

Ancient criteria

'Ancient criteria' are for example pathovars, phenotypically and biochemically based criteria used to distinguish between *E. coli* and *Shigella* spp. before genomic criteria. A first genomic criterion was G+C content. Based on GC% comparison between strains, it can be classified as the same species or not [3]. Variation is lower than 2% within *E. coli* (50.4–51.2) and including *Shigella* (50.4–51.2). Variation is lower than 1% within *Shigella* (50.7–51.1).

Shigella spp. are indistinguishable from *E. coli* by DNA/DNA hybridization [69]. In the 16S identity matrix comparing all strains, we noticed that the lowest identity was 98.83%. The minimal 16S identity within *E. coli* was 99.41% between *E. coli* IHE3034 and *E. coli* UMNK88, whereas the minimal identity between *E. coli* and *Shigella* was 99.03% between *E. coli* O26 H11 11368 and *Shigella dysenteriae* Sd197. The identity between *E. coli* and *Shigella* spp. exceeds the cut-offs used to classify bacterial isolates at the genus and species levels on the basis of 16S rRNA gene sequence identity values (95 and 97% or 98.7%, respectively). In general, *Shigella* and *E. coli* appear to belong to the same species and some *Shigella* were closer to some *E. coli* than to some other *Shigella*.

New pangenomic criteria

To use pangenome for taxonomy, we clustered our genomes based on COG and KEGG data. In both cases, *Shigella* was included inside the *E. coli* cluster and did not constitute a separate group. Then, we looked at the phylogenetic tree based on the concatenation of the core gene SNPs (not shown); *Shigella* did not constitute a unique cluster, instead the species tended to be distributed among the different *E. coli* clusters. Then, we calculated the distance between genomes on the basis of nucleic sequence identity, which revealed that some *E. coli*

(26 out of 42 genomes) were closer to *Shigella* (with more than 90% similarity) than to some other *E. coli* (with around 80% similarity). The principal coordinate analysis, based on the nucleotide similarity between genomes, showed several different clusters including two clusters containing a mix of *Shigella* and *E. coli* species.

Pangenome and taxonomy

Thanks to USEARCH [70] for protein de-replication, followed by a tBLASTn with a 10E-3 E-value, we determined the core/pangenome ratio, the pangenome and the core genome values after each added strain for *E. coli*, *E. coli* + *Shigella*, *Shigella* and *Shigella flexneri* (Fig. 2). For each curve (core, pangenome and ratio), we looked for the best R^2 (coefficient of determination) in order to determine the most accurate regression type. We also calculated the average rift between the core and the pangenome curves.

In all cases, the pangenome curve is described as a linear function whereas those from the core and the ratio are described by power functions. When it is a single species, like *S. flexneri* (Fig. 2a), core, pangenome and ratio curves matched perfectly with their trend curves corresponding to their function. First, in Fig. 2(b), the ratio and the pangenome curves showed that there are different species, because at some points curves did not follow the trend curve. Then, in Fig. 2(c), the addition of nine *Shigella* to the 42 *E. coli* samples creates a break in the pangenome and ratio curves. This is in correlation with the disappearance of 543 functions, 216 from *E. coli* and 327 from *Shigella*. Indeed, the standard deviation between the core curve and the pangenome curve has only a 1% variation between the two conditions (with or without *Shigella*).

Finally, with the addition of a second *E. coli* we can see there is a great decrease (15%) in the ratio, whereas in a homogeneous species, like *S. flexneri*, this decrease is only 2%.

In conclusion, we focused on the fact that *E. coli* is not a homogeneous species, with these variations between trend curves and ratio (or pangenome curve), compared with *S. flexneri*, which is a homogeneous species. There is also a breakpoint in the ratio and pangenome curves. Mathematically, this corresponds to the start of a new function. Here, this points to the start of a new species, which may be explored further as a new species criterion to define species.

Relations Between Pangenome and Lifestyle

Ratio

Finally, based on the 'backbone files' [44] of MAUVE, we calculated the size of the core and the pangenome of 27 species (Table 3). After determining the core/pangenome ratio (Table 3), we noticed that the species with a closed pangenome possessed a ratio $\geq$89% and that they were all allopatric. For instance, the species raising the smallest ratio (5%) was a sympatric bacterium that lived in a marine environment. This ratio is based on both coding region and intergenic regions. We also calculated a ratio only with the coding part, based on the core genes.

TABLE 3. Ratio core/pangenome of several bacterial species according to their life style

Species	Genome used	Lifestyle	Intracellular	Niche	%[a]
Prochlorococcus marinus	12	Sympatric	no	Marine environment	5
Clostridium botulinum	14	Sympatric	no	Soil	11
Rhodopseudomonas palustris	7	Sympatric	no	Soil, marine environment	46
Sinorhizobium meliloti	6	Sympatric	no	Soil	49
Salmonella enterica	20	Sympatric	facultative	Animals	62
Acinetobacter baumannii	11	Sympatric	no	?	65
Legionella pneumophila	11	Sympatric	facultative	Amoeba	69
Escherichia coli	19	Sympatric	no	Animals	70
Bacillus cereus	12	Sympatric	no	Soil	74
Campylobacter jejuni	14	Sympatric	facultative	Human, chicken	76
Clostridium difficile	18	Sympatric	no	Human gut	77
Helicobacter pylori	10	Sympatric	facultative	Human	78
Haemophilus influenzae	9	Sympatric	facultative	Human	80
Streptococcus pneumoniae	10	Sympatric	no	Human	82
Pseudomonas aeruginosa	7	Sympatric	no	Water	84
Streptococcus agalactiae	5	Sympatric	no	Human	84
Listeria monocytogenes	20	Sympatric	facultative	Amoeba?	84
Francisella tularensis	13	Sympatric	facultative	Ticks	87
Yersinia pestis	12	Allopatric	facultative	Rodents	89
Coxiella burnetii	7	Allopatric	yes	Animals	90
Tropheryma whipplei	19	Allopatric	yes	Human	94
Mycobacterium tuberculosis	20	Allopatric	yes	Human	96
Buchnera aphidicola	8	Allopatric	yes	Aphid	98
Bacillus anthracis	9	Allopatric	no	Animals	99
Rickettsia rickettsii	8	Allopatric	yes	Ticks	99
Chlamydia trachomatis	20	Allopatric	yes	Human	99
Rickettsia prowazekii	8	Allopatric	yes	Human	100

[a]% is the ratio core/pangenome.

Pangenome size and lifestyle

The size of a pangenome is strongly related to the balance existing between gene gain and loss events (Fig. 3). When an ecosystem becomes different (Fig. 3), some functions can then become useless and eventually be lost. In contrast, when the bacteria are in a very diverse environment with many partners, gain events are common (Fig. 3). The genome size is also strongly connected to the selfish genes, which are parasitic and constitute the mobilome (see above). Phages, integrases and transposases contribute to the increase in genome size and are the consequence of life in community. Usually, the more partners there are, the greater the probability of acquiring parasitic DNA. A sympatric bacterium will then have a wide and open pangenome and will possess a quite consequent mobilome as well as more defence mechanisms (CRISPRs) than intracellular and allopatric species, which will have a small and closed pangenome [44].

Case of 'bad bugs'

It is known that intracellular bacteria possess fewer genes for transcription [71] and there is a decrease of genes involved in metabolism [72]. In 2011, a study of 'bad bugs' (targeting the 12 most dangerous bacteria for human beings) [59] was conducted. Globally, it was noticed that the virulent isolates tend to have a reduced genome compared with their commensal counterparts, but above all that there are functional reductions. Indeed, of the 23 tested COG categories [59], a decrease in gene number was found in ten, specifically for transcription and amino acid metabolism. It was noted that the genes lost from the 'bad bugs' mainly encode for the metabolism and transport functions.

Pangenome and Lifestyle Examples, *Yersinia pestis* and *Bacillus anthracis*

Yersinia pestis [73], the plague's agent, was studied in 2010. After sequencing 14 genomes, assembly was carried out using Celera assembler [74] and annotation using the MANATEE tool (http://manatee.sourceforge.net/). After global alignment of genomes using MUMmer [75], pangenome composition was predicted using WU-BLASTp and tBLASTn. The core genome consists of 3668 genes and, as for every closed pangenome, the addition of new isolates changed almost nothing.

Although *Y. pestis* lived in the soil, it had a closed pangenome reflecting an allopatric lifestyle. This was the same as *B. anthracis*, which lived in a dormant form in the soil and which multiplied in its host. Hence, the pangenome makes it possible to determine if a bacteria is just resting and not multiplying in an environment with many other microorganisms (such as soil or water) or if it is active. Take *B. anthracis* for instance, which lives in the soil in a dormant sporulated form. When it becomes active and multiplies in its host, it has few chances to exchange genes. Therefore the *B. anthracis* pangenome is closed with a core/pangenome ratio of 99%.

Conclusion: a Quantic Perspective for Taxonomy of Pathogenic Species

Pangenome studies have become almost essential for bacterial genome comparisons. After carefully choosing the strains of interest, we can select an experimental method such as a

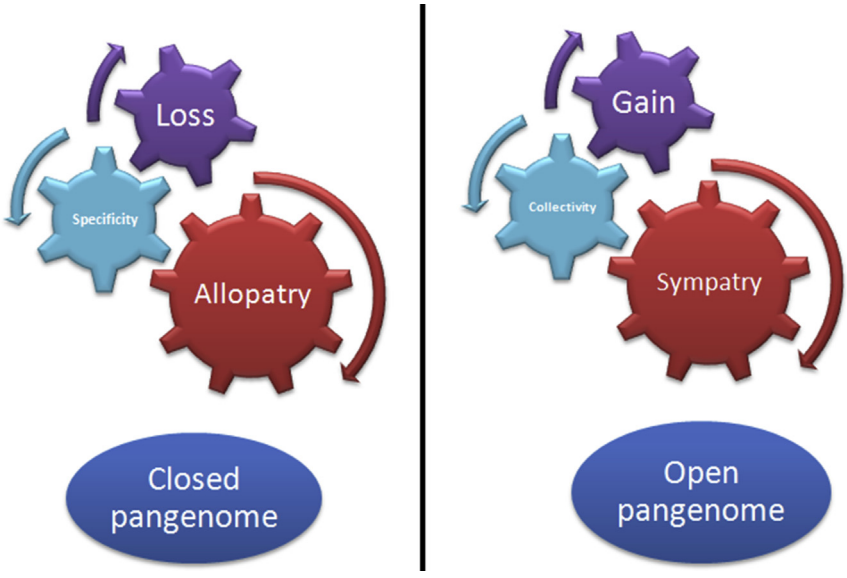

FIG. 3. Summary of the difference between closed and open pangenome.

microarray [26] or bioinformatics-based method (Fig. 4). Bioinformatics offers tools serving general [37] and dedicated [53,54] purposes. Thanks to these analyses, study of pangenomes can provide different kinds of data and increase our knowledge and understanding of a species.

First, the size of a genome is directly correlated to its capacity to acquire, or not, exogenous DNA, to gain and loss events and to the presence of selfish genes. The pangenome size depends on all these parameters. Hence, depending on its size and on its type (open or closed), we can determine the species' lifestyle (allopatry or sympatry), and also have an idea of the number of genomes we need to have the best view of real genomic content (Fig. 3).

Pangenome study also allowed us to find the resistome [47], the non-virulence genes [62] and the mobilome [48] (to determine selfish genes) of a group of strains(Fig. 4). Sometimes it is possible to extrapolate the age of clones by studying SNP in the core genome. Moreover, by analogy with the pangenome concept, the panmetabolism can be described, giving a large but detailed view of all common metabolisms and/or differences in the strains of interest.

By grouping all these genomic data and the lifestyle information, it may be possible to redefine species and classify them depending on their genomic content. Indeed, groups of strains with a core/pangenome ratio of 100 or 99%, with a very reduced mobilome and with an identical gene content may be considered to be one species. However, in the case of an infinite pangenome, such as *E. coli*, or in the case of a very small ratio (5%) like *Prochlorococcus marinus*, can we talk about species yet? Instead of a single species, do we have a complex of species? Definitions of species were often reached using old tools. Moreover, some species are, by nature, non-homogeneous (in the case of sympatric species). So redefining species [76] may be an interesting perspective for the future, using a combination of pangenomic data, phylogeny and phylogenomics (unpublished data).

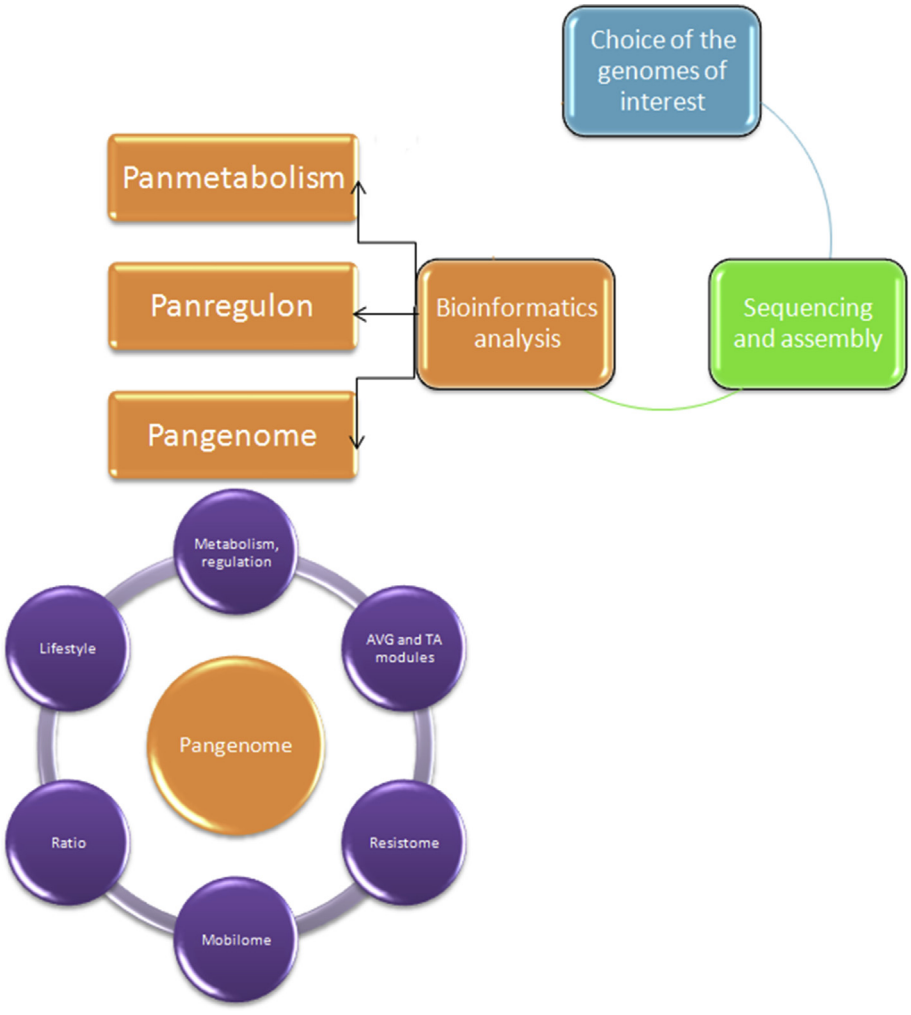

FIG. 4. Strategy of analyses of the pangenome.

Besides redefining species, the second important key to the study of the pangenomes is to see what is not visible at first glance. Take *B. anthracis* for example, which lives in a niche appearing sympatric (the soil) [77] but remains dormant in spore form and has a very closed pangenome. Conversely, *L. pneumophila* is intracellular, but it is metabolically active in its niche (amoeba) [68,78] and has an open pangenome. The pangenome therefore also provides an alternative method for analysing lifestyle, which is not simply looking at the apparent predicted niche.

Future Perspectives

In terms of future perspectives, we can consider applying the pangenome to the reclassification of other bacterial pathogenic species or genus, such as *Salmonella*.

Conflict of interest

None declared.

Appendix A. Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.nmni.2015.06.005.

References

[1] Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 2008;11:472–7.
[2] Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005;15:589–94.
[3] Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M, Sentausa E, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol 2014;64:384–91.
[4] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 2005;102:13950–5.
[5] Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, et al. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 2007;8:R103.
[6] Boissy R, Ahmed A, Janto B, Earl J, Hall BG, Hogg JS, et al. Comparative supragenomic analyses among the pathogens *Staphylococcus aureus, Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics 2011;12:187.
[7] Georgiades K, Raoult D. Defining pathogenic bacterial species in the genomic era. Front Microbiol 2010;1:151.
[8] Snipen L, Almoy T, Ussery DW. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 2009;10:385.
[9] Lefebure T, Bitar PD, Suzuki H, Stanhope MJ. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol Evol 2010;2:646–55.
[10] Diene SM, Merhej V, Henry M, El Filali A, Roux V, Robert C, et al. The rhizome of the multidrug-resistant Enterobacter aerogenes genome reveals how new "killer bugs" are created because of a sympatric lifestyle. Mol Biol Evol 2013 Feb;30(2):369–83.
[11] Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 2006;361:1929–40.
[12] Staley JT. Universal species concept: pipe dream or a step toward unifying biology? J Ind Microbiol Biotechnol 2009;36:1331–6.
[13] Doolittle WF, Papke RT. Genomics and the bacterial species problem. Genome Biol 2006;7:116.
[14] Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, et al. Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol 2005;3:733–9.
[15] Dagan T, Martin W. The tree of one percent. Genome Biol 2006;7:118.
[16] Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 2008;190:6881–93.
[17] Amit U, Porat N, Basmaci R, Bidet P, Bonacorsi S, Dagan R, et al. Genotyping of invasive *Kingella kingae* isolates reveals predominant clones and association with specific clinical syndromes. Clin Infect Dis 2012;55:1074–9.
[18] Xiong X, Wang X, Wen B, Graves S, Stenos J. Potential serodiagnostic markers for Q fever identified in *Coxiella burnetii* by immunoproteomic and protein microarray approaches. BMC Microbiol 2012;12:35.
[19] Arricau-Bouvery N, Hauck Y, Bejaoui A, Frangoulidis D, Bodier CC, Souriau A, et al. Molecular characterization of *Coxiella burnetii* isolates by infrequent restriction site-PCR and MLVA typing. BMC Microbiol 2006;6:38.
[20] Roest HI, Ruuls RC, Tilburg JJ, Nabuurs-Franssen MH, Klaassen CH, Vellema P, et al. Molecular epidemiology of *Coxiella burnetii* from ruminants in Q fever outbreak, the Netherlands. Emerg Infect Dis 2011;17:668–75.
[21] Tilburg JJ, Rossen JW, van Hannen EJ, Melchers WJ, Hermans MH, van de Bovenkamp J, et al. Genotypic diversity of *Coxiella burnetii* in the 2007–2010 Q fever outbreak episodes in The Netherlands. J Clin Microbiol 2012;50:1076–8.
[22] Reuter S, Harrison TG, Koser CU, Ellington MJ, Smith GP, Parkhill J, et al. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open 2013;3.
[23] Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc Natl Acad Sci U S A 2009;106:15442–7.
[24] Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 2011;365:709–17. http://dx.doi.org/10.1056/NEJMoa1106920.
[25] Feng S, Tillier ER. A fast and flexible approach to oligonucleotide probe design for genomes and gene families. Bioinformatics 2007;23:1195–202.
[26] Hug LA, Salehi M, Nuin P, Tillier ER, Edwards EA. Design and verification of a pangenome microarray oligonucleotide probe set for *Dehalococcoides* spp. Appl Environ Microbiol 2011;77:5361–9.

[27] Phillippy AM, Deng X, Zhang W, Salzberg SL. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 2009;10:293.
[28] Leroy Q, Armougom F, Barbry P, Raoult D. Genomotyping of *Coxiella burnetii* using microarrays reveals a conserved genomotype for hard tick isolates. PLoS One 2011;6:e25781.
[29] Leroy Q, Raoult D. Review of microarray studies for host-intracellular pathogen interactions. J Microbiol Methods 2010;81:81–95.
[30] Bayjanov JR, Wels M, Starrenburg M, van Hylckama Vlieg JE, Siezen RJ, et al. PanCGH: a genotype-calling algorithm for pangenome CGH data. Bioinformatics 2009;25:309–14.
[31] Bayjanov JR, Siezen RJ, van Hijum SA. PanCGHweb: a web tool for genotype calling in pangenome CGH data. Bioinformatics 2010;26:1256–7.
[32] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[33] Chen F, Mackey AJ, Stoeckert Jr CJ, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 2006;34:D363–8.
[34] Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, et al. Unity in variety—the pan-genome of the Chlamydiae. Mol Biol Evol 2011;28:3253–70.
[35] D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics 2010;11:181.
[36] Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C. The *Salmonella enterica* pan-genome. Microb Ecol 2011;62:487–504.
[37] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–8.
[38] Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 2001;29:37–40.
[39] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34.
[40] Barakat M, Ortet P, Whitworth DE. P2RP: a web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes. BMC Genomics 2013;14:269.
[41] Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5:e11147.
[42] Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011;28:2731–9.
[43] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–66.
[44] Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, et al. Progressive genome-wide introgression in agricultural *Campylobacter coli*. Mol Ecol 2013;22:1051–64.
[45] Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, et al. Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. Proc Natl Acad Sci U S A 2013;110:577–82.
[46] Liu B, Pop M. ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res 2009;37:D443–7.
[47] Joshi RS, Jamdhade MD, Sonawane MS, Giri AP. Resistome analysis of *Mycobacterium tuberculosis*: identification of aminoglycoside 2'-Nacetyltransferase (AAC) as co-target for drug desigining. Bioinformation 2013;9:174–81.
[48] den Bakker HC, Cummings CA, Ferreira V, Vatta P, Orsi RH, Degoricija L, et al. Comparative genomics of the bacterial genus *Listeria*: genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 2010;11:688.
[49] Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007;35:W52–7.
[50] Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res 2011;39:W347–52.
[51] Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75.
[52] Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 2006;34:D32–6.
[53] Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 2012;40:e172.
[54] Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. Bioinformatics 2012;28:416–8.
[55] Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 2010;11:461.
[56] Unterholzner SJ, Poppenberger B, Rozhon W. Toxin-antitoxin systems: biology, identification, and application. Mob Genet Elements 2013;3:e26219.
[57] Socolovschi C, Audoly G, Raoult D. Connection of toxin-antitoxin modules to inoculation eschar and arthropod vertical transmission in Rickettsiales. Comp Immunol Microbiol Infect Dis 2013;36:199–209.
[58] Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. Brief Funct Genomics 2013;12:291–304.
[59] Georgiades K, Raoult D. Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. PLoS One 2011;6:e17962.
[60] Georgiades K, Raoult D. Comparative genomics evidence that only protein toxins are tagging bad bugs. Front Cell Infect Microbiol 2011;1:7.
[61] Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, et al. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 2001;293:2093–8.
[62] Bliven KA, Maurelli AT. Antivirulence genes: insights into pathogen evolution through gene loss. Infect Immun 2012;80:4061–70.
[63] Olivares J, Bernardini A, Garcia-Leon G, Corona F, Sanchez B, Martinez JL. The intrinsic resistome of bacterial pathogens. Front Microbiol 2013;4:103.
[64] Walsh F, Duffy B. The culturable soil antibiotic resistome: a community of multi-drug resistant bacteria. PLoS One 2013;8:e65567.
[65] Vieira G, Sabarly V, Bourguignon PY, Durot M, Le FF, Mornico D, et al. Core and panmetabolism in *Escherichia coli*. J Bacteriol 2011;193:1461–72.
[66] Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, Lucas S, et al. Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. BMC Genomics 2011;12:235.
[67] Oliver HF, Orsi RH, Wiedmann M, Boor KJ. *Listeria monocytogenes* σ^B has a small core regulon and a conserved role in virulence but makes differential contributions to stress tolerance across a diverse collection of strains. Appl Environ Microbiol 2010;76:4216–32.
[68] Gimenez G, Bertelli C, Moliner C, Robert C, Raoult D, Fournier PE, et al. Insight into cross-talk between intra-amoebal pathogens. BMC Genomics 2011;12:542.

[69] Welch Rodney A. The Genus *Escherichia*. In: Dworkin Martin, Falkow Stanley, Rosenberg Eugene, editors. The Prokaryotes, vol. 6. New York: Springer-Verlag; 2006. p. 60–71 [Chapter 3.3.3].
[70] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–1.
[71] Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 2009;4:13.
[72] Georgiades K, Merhej V, El KK, Raoult D, Pontarotti P. Gene gain and loss events in *Rickettsia* and *Orientia* species. Biol Direct 2011;6:6.
[73] Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, et al. Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. J Bacteriol 2010;192:1685–99.
[74] Huson DH, Reinert K, Kravitz SA, Remington KA, Delcher AL, Dew IM, et al. Design of a compartmentalized shotgun assembler for the human genome. Bioinformatics 2001;17(Suppl. 1):S132–9.
[75] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5:R12.
[76] Scortichini M, Marcelletti S, Ferrante P, Firrao G. A genomic redefinition of species. PLoS One 2013;8:e75794.
[77] Wang DB, Tian B, Zhang ZP, Deng JY, Cui ZQ, Yang RF, et al. Rapid detection of Bacillus anthracis spores using a super-paramagnetic lateral-flow immunological detection system. Biosens Bioelectron 2012. http://dx.doi.org/10.1016/j.bios.2012.10.088. pii: S0956-5663(12)00782-8.
[78] Moliner C, Fournier PE, Raoult D. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol Rev 2010;34:281–94.
[79] Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 2010;11:R107.
[80] Gerrish RS, Gill AL, Fowler VG, Gill SR. Development of pooled suppression subtractive hybridization to analyze the pangenome of *Staphylococcus aureus*. J Microbiol Methods 2010;81:56–60.
[81] Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, Yamaoka Y, et al. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. PLoS Genet 2005;1:e43.
[82] Thompson CC, Vicente AC, Souza RC, Vasconcelos AT, Vesth T, Alves Jr N, et al. Genomic taxonomy of *Vibrios*. BMC Evol Biol 2009;9:258.
[83] Zakham F, Belayachi L, Ussery D, Akrim M, Benjouad A, El AR, et al. Mycobacterial species as case-study of comparative genome analysis. Cell Mol Biol (Noisy-le-grand) 2011;57(Suppl.):OL1462–9.
[84] Imperi F, Antunes LC, Blom J, Villa L, Iacono M, Visca P, et al. The genomics of *Acinetobacter baumannii*: insights into genome plasticity, antimicrobial resistance and pathogenicity. IUBMB Life 2011;63:1068–74.
[85] Deng X, Phillippy AM, Li Z, Salzberg SL, Zhang W. Probing the pangenome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. BMC Genomics 2010;11:500.
[86] Kuenne C, Billion A, Mraheil MA, Strittmatter A, Daniel R, Goesmann A, et al. Reassessment of the *Listeria monocytogenes* pangenome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. BMC Genomics 2013;14:47.
[87] Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tummler B. *Pseudomonas aeruginosa* genomic structure and diversity. Front Microbiol 2011;2:150.
[88] van SW, Top J, Riley DR, Boekhorst J, Vrijenhoek JE, Schapendonk CM, et al. Pyrosequencing-based comparative genome analysis of the nosocomial pathogen *Enterococcus faecium* and identification of a large transferable pathogenicity island. BMC Genomics 2010;11:239.
[89] Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF. Analysis of ultra low genome conservation in *Clostridium difficile*. PLoS One 2010;5:e15147.
[90] Wilson MK, Lane AB, Law BF, Miller WG, Joens LA, Konkel ME, et al. Analysis of the pan genome of *Campylobacter jejuni* isolates recovered from poultry by pulsed-field gel electrophoresis, multilocus sequence typing (MLST), and repetitive sequence polymerase chain reaction (rep-PCR) reveals different discriminatory capabilities. Microb Ecol 2009;58:843–55.
[91] Mira A, Martin-Cuadrado AB, D'Auria G, Rodriguez-Valera F. The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol 2010;13:45–57.
[92] Xu Z, Chen X, Li L, Li T, Wang S, Chen H, et al. Comparative genomic characterization of *Actinobacillus pleuropneumoniae*. J Bacteriol 2010;192:5625–36.
[93] Lefebure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol 2007;8:R71.
[94] Zhang A, Yang M, Hu P, Wu J, Chen B, Hua Y, et al. Comparative genomic analysis of *Streptococcus suis* reveals significant genomic diversity among different serotypes. BMC Genomics 2011;12:523.
[95] Kittichotirat W, Bumgarner RE, Asikainen S, Chen C. Identification of the pangenome and its components in 14 distinct *Aggregatibacter actinomycetemcomitans* strains by comparative genomic analysis. PLoS One 2011;6:e22420.
[96] Barrangou R, Briczinski EP, Traeger LL, Loquasto JR, Richards M, Horvath P, et al. Comparison of the complete genome sequences of *Bifidobacterium animalis* subsp. *lactis* DSM 10140 and BI-04. J Bacteriol 2009;191:4144–51.
[97] Remenant B, Coupat-Goutaland B, Guidot A, Cellier G, Wicker E, Allen C, et al. Genomes of three tomato pathogens within the *Ralstonia solanacearum* species complex reveal significant evolutionary divergence. BMC Genomics 2010;11:379.
[98] Mann RA, Smits TH, Buhlmann A, Blom J, Goesmann A, Frey JE, et al. Comparative genomics of 12 strains of *Erwinia amylovora* identifies a pan-genome with a large conserved core. PLoS One 2013;8:e55644.
[99] Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, et al. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One 2013;8:e53818.
[100] Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, et al. Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 2012;13:533.
[101] Liang W, Zhao Y, Chen C, Cui X, Yu J, Xiao J, et al. Pan-genomic analysis provides insights into the genomic variation and evolution of *Salmonella* Paratyphi A. PLoS One 2012;7:e45346.
[102] Borneman AR, McCarthy JM, Chambers PJ, Bartowsky EJ. Comparative analysis of the *Oenococcus oeni* pan genome reveals genetic diversity in industrially-relevant pathways. BMC Genomics 2012;13:373.
[103] Conlan S, Mijares LA, Becker J, Blakesley RW, Bouffard GG, Brooks S, et al. *Staphylococcus epidermidis* pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol 2012;13:R64.
[104] Trost E, Blom J, Soares SC, Huang IH, Al-Dilaimi A, Schroder J, et al. Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol 2012;194:3199–215.
[105] Silby MW, Cerdeno-Tarraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM, et al. Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol 2009;10:R51.

[106] Ho CC, Lau CC, Martelli P, Chan SY, Tse CW, Wu AK, et al. Novel pan-genomic analysis approach in target selection for multiplex PCR identification and detection of *Burkholderia pseudomallei, Burkholderia thailandensis*, and *Burkholderia cepacia* complex species: a proof-of-concept study. J Clin Microbiol 2009;49:814–21.

[107] Ussery DW, Kiil K, Lagesen K, Sicheritz-Ponten T, Bohlin J, Wassenaar TM. The genus *Burkholderia*: analysis of 56 genomic sequences. Genome Dyn 2009;6:140–57.

[108] Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, et al. Comparative genomics of the genus *Bifidobacterium*. Microbiology 2010;156:3243–54.

[109] Ahsanul IM, Edwards EA, Mahadevan R. Characterizing the metabolism of *Dehalococcoides* with a constraint-based model. PLoS Comput Biol 2010;6.

[110] Liu W, Fang L, Li M, Li S, Guo S, Luo R, et al. Comparative genomics of *Mycoplasma*: analysis of conserved essential genes and diversity of the pan-genome. PLoS One 2012;7:e35698.

[111] Blumer-Schuette SE, Giannone RJ, Zurawski JV, Ozdemir I, Ma Q, Yin Y, et al. Caldicellulosiruptor core and pangenomes reveal determinants for noncellulosomal thermophilic deconstruction of plant biomass. J Bacteriol 2012;194:4015–28.