www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Genomic characterization and computational phenotyping of nitrogen-fixing bacteria isolated from Colombian sugarcane fields

Luz K. Medina-Cordoba[1,2], Aroon T. Chande[1,2,3], Lavanya Rishishwar[1,2,3], Leonard W. Mayer[2,3], Lina C. Valderrama-Aguirre[2,4], Augusto Valderrama-Aguirre[2,5], John Christian Gaby[1], Joel E. Kostka[1,2,7]✉ & I. King Jordan[1,2,3,6]✉

Previous studies have shown the sugarcane microbiome harbors diverse plant growth promoting microorganisms, including nitrogen-fixing bacteria (diazotrophs), which can serve as biofertilizers. The genomes of 22 diazotrophs from Colombian sugarcane fields were sequenced to investigate potential biofertilizers. A genome-enabled computational phenotyping approach was developed to prioritize sugarcane associated diazotrophs according to their potential as biofertilizers. This method selects isolates that have potential for nitrogen fixation and other plant growth promoting (PGP) phenotypes while showing low risk for virulence and antibiotic resistance. Intact nitrogenase (*nif*) genes and operons were found in 18 of the isolates. Isolates also encode phosphate solubilization and siderophore production operons, and other PGP genes. The majority of sugarcane isolates showed uniformly low predicted virulence and antibiotic resistance compared to clinical isolates. Six strains with the highest overall genotype scores were experimentally evaluated for nitrogen fixation, phosphate solubilization, and the production of siderophores, gibberellic acid, and indole acetic acid. Results from the biochemical assays were consistent and validated computational phenotype predictions. A genotypic and phenotypic threshold was observed that separated strains by their potential for PGP versus predicted pathogenicity. Our results indicate that computational phenotyping is a promising tool for the assessment of bacteria detected in agricultural ecosystems.

The human population is expected to increase 45% by the year 2050, which will in turn lead to a massive increase in the global demand for food[1]. Given the scarcity of arable land worldwide, an increase in agricultural production of this magnitude will require vast increases in cropping intensity and yield[2]. It has been estimated that as much as 90% of the increase in global crop production will need to come from increased yield alone[3]. At the same time, climate change and other environmental challenges will necessitate the development of agricultural practices that are more ecologically friendly and sustainable.

Chemical fertilizers that provide critical macronutrients to crops—such as nitrogen (N), phosphorus (P), potassium (K), and sulfur (S)—are widely used to maximize agricultural yield[4]. The application of chemical fertilizers represents a major cost for agricultural companies and also contributes to environmental damage, such as air pollution through the formation of microparticles, soil depletion, and water contamination via run-off[5]. Biological fertilizers (biofertilizers) are comprised of microbial inoculants that promote plant growth, thereby representing an alternative or complementary approach for increasing crop yield, which is more sustainable and environmentally friendly. Biofertilizers augment plant growth through nutrient acquisition, hormone production, and by boosting immunity to pathogens[6].

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [2]PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia. [3]Applied Bioinformatics Laboratory, Atlanta, GA, USA. [4]Laboratory of Microorganismal Production (Bioinoculums), Department of Field Research in Sugarcane, INCAUCA S.A.S., Cali, Valle del Cauca, Colombia. [5]Universidad Santiago de Cali, Cali, Colombia. [6]Present address: School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Dr NW, Atlanta, GA 30332, USA. [7]Present address: School of Biological Sciences, Georgia Institute of Technology, 310 Ferst Dr NW, Atlanta, GA 30332, USA. ✉email: joel.kostka@biology.gatech.edu; king.jordan@biology.gatech.edu

Sample ID	Genome Length (bp)	N50[a]	L50[b]	GC (%)	# of Contigs[c]
SCK1	4,522,541	402,304	4	66.79	24
SCK2	5,231,439	417,927	5	59.33	53
SCK3	3,824,428	670,745	3	41.82	150
SCK4	4,511,030	223,239	8	66.79	55
SCK5	5,774,634	162,673	13	53.1	98
SCK6	6,094,823	117,689	15	56.73	294
SCK7	5,693,007	282,996	7	57.03	50
SCK8	5,695,902	281,292	9	57.03	50
SCK9	5,579,618	311,650	6	57.03	42
SCK10	5,591,472	614,324	3	57.03	34
SCK11	5,696,136	382,597	5	57.15	268
SCK12	5,817,089	176,655	10	57.02	79
SCK13	5,476,221	358,490	5	57.34	33
SCK14	5,465,811	300,899	5	57.34	41
SCK15	5,564,330	330,579	5	57.15	43
SCK16	5,795,921	478,592	3	54.06	84
SCK17	5,475,984	358,490	4	57.34	35
SCK18	5,476,135	422,400	3	57.34	32
SCK19	5,688,396	270,585	7	57.09	56
SCK20	5,500,801	82,111	20	57.45	165
SCK21	5,324,920	112,078	15	55.26	100
SCK22	5,847,607	65,329	29	57.02	181

Table 1. Genome assembly statistics for the isolates characterized here. [a]When the contigs of an assembly are arranged from largest to smallest, N50 is the length of the contig that makes up at least 50% of the genome. [b]L50 is the number of contigs equal to or longer than N50. [c]Number of contigs ≥ 500 bp in length.

Sugarcane is a tall, perennial grass cultivated in tropical and warm temperate regions around the world, which is capable of producing high concentrations of sugar (sucrose) and generating diverse byproducts[7]. Sugarcane is consistently ranked as one of the top ten planted crops in the world[8]. Sugarcane agriculture plays a vital role in the economy of Colombia by supporting the production of food, energy, and fuel (ethanol) along with a variety of organic by-products. Our group is working to help develop more effective and sustainable sugarcane cropping practices in Colombia. The long-term goals of this work are to simultaneously (i) increase crop yield, and (ii) decrease the reliance on chemical fertilizers via the discovery, characterization, and application of endemic (native) biofertilizers to Colombian sugarcane fields.

Most sugarcane companies in Colombia currently use commercially available biofertilizers, consisting primarily of nitrogen-fixing bacteria, which were discovered and isolated from other countries (primarily Brazil), with limited success. We hypothesized that indigenous bacteria should be better adapted to the local environment and thereby serve as more effective biofertilizers for Colombian sugarcane. The use of indigenous bacteria as biofertilizers should also mitigate potential threats to the environment posed by non-native, and potentially invasive, species of bacteria. Finally, indigenous bacteria represent a renewable resource that agronomists can continually develop through isolation and cultivation of local strains.

The advent of next-generation sequencing technologies has catalyzed the development of genome-enabled approaches to harness plant microbiomes in sustainable agriculture[9,10]. The objective of this study was to use genome analysis to predict the local bacterial isolates that have the greatest potential for plant growth promotion while representing the lowest risk for virulence and antibiotic resistance. Putative biofertilizer strains were isolated and cultivated from Colombian sugarcane fields, and computational phenotyping was employed to predict their potential utility of strains as biofertilizers. We then performed a laboratory evaluation of predict the potential utility of these strains as biofertilizers, with the aim of validating our computational phenotyping approach.

Results

Initial genome characterization of putative nitrogen-fixing bacteria.

A systematic cultivation approach, incorporating seven carbon substrates in nitrogen-free media (Supplementary Figure S1), was employed to isolate putative nitrogen-fixing bacteria from the four different sugarcane plant compartments, and isolates were screened for nitrogen fixation potential through PCR amplification of *nifH* genes. This initial screening procedure yielded several hundred clonal isolates of putative nitrogen-fixing bacteria, and Ribosomal Intergenic Spacer Analysis (RISA) was subsequently used to identify the (presumably) genetically unique strains from the larger set of clonal isolates. A total of 22 potentially unique strains of putative nitrogen-fixing bacteria were isolated in this way and selected for genome sequence analysis.

Genome sequencing and assembly summary statistics for the 22 isolates are shown in Table 1. Isolate genomes were sequenced to an average of 67x coverage (range 50x—88x) and genome sizes range from 4.5 to 6.1 Mb. GC content varies from 41.82 to 66.69%, with a distinct mode at ~ 57%. The genome assemblies show a range of

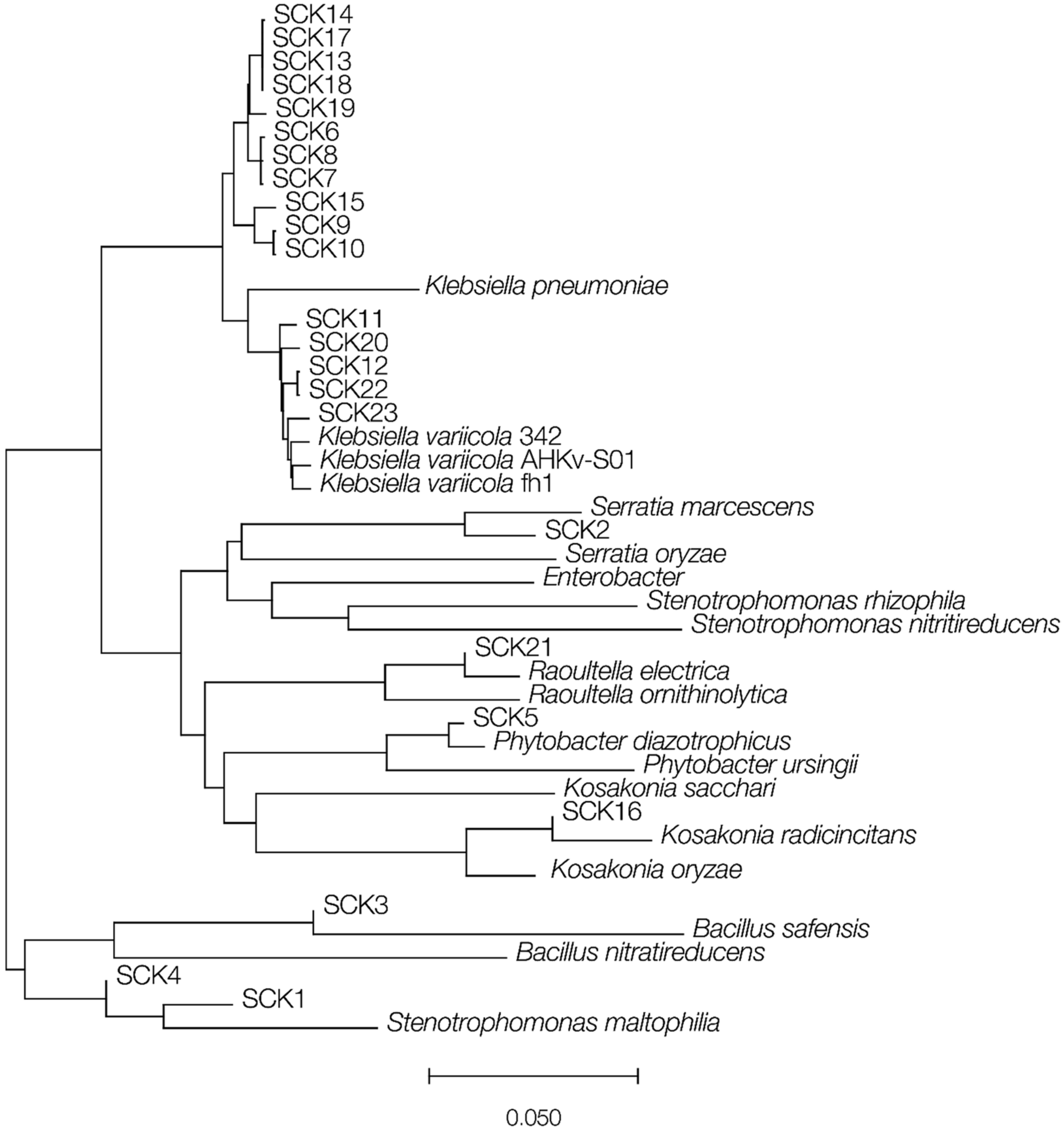

Figure 1. Phylogeny of the bacterial isolates characterized here (SCK numbers) together with their most closely related bacterial type strains. The phylogeny was reconstructed using pairwise average nucleotide identities between whole genome sequence assemblies, converted to p-distances, with the neighbor-joining method. Horizontal branch lengths are scaled according the p-distances as shown.

24–294 contigs ≥ 500 bp in length, with N50 values that range from 65,329 to 670,745bp (avg. = 310,166bp) and L50 values that range from 3 to 29 (avg. = 8.4). Genome sequence assemblies, along with their functional annotations, can all be found using the NCBI BioProject PRJNA418312. Individual BioSample, Genbank Accession, and Assembly Accession numbers for the 22 isolates are shown in Supplementary Table S1.

Comparative genomic analysis. Average nucleotide identity (ANI; Fig. 1) and 16S rRNA sequence analysis (Supplementary Figure S2) were used to assign the species (genus) origins for the 22 putative nitrogen-fixing isolate genome sequences and the results of both approaches are highly concordant (Table 2), with ANI yielding superior resolution to 16S rRNA sequence analysis. A total of seven different species and seven different genera were identified among the 22 isolates characterized here. Analysis of *nifH* gene sequences also gave similar results; however, four of the isolates were not found to encode *nifH* genes, despite their (apparent) ability to grow on nitrogen-free media and the positive *nifH* PCR results. As described in the "Methods" section, we used the Rapid Annotations using Subsystems Technology (RAST) server to predict and annotate genes from our assemblies, and this approach was unable to detect *nifH* genes in those 4 isolates. We performed additional analyses on these four genomes to confirm the absence of *nifH*: we used NCBI BLAST with *nifH* nucleotide and amino acid queries to search the genomes and the *nifH*-specific tool TaxADivA[11]. Neither of these additional analyses found

Strain	ANI	16S	*nifH*[a]
SCK1	*Stenotrophomonas sp.*	*Stenotrophomonas*	NA
SCK2	*Serratia marcescens*	*Serratia*	NA
SCK3	*Bacillus safensis*	*Bacillus*	NA
SCK4	*Stenotrophomonas sp.*	*Stenotrophomonas*	NA
SCK5	*Phytobacter diazotrophicus*	*Phytobacter*	*Phytobacter*
SCK6	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK7	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK8	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK9	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK10	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK11	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK12	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK13	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK14	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK15	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK16	*Kosakonia radicincitans*	*Kosakonia*	*Kosakonia*
SCK17	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK18	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK19	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK20	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*
SCK21	*Raoultella electrica*	*Raoultella*	*Raoultella*
SCK22	*Klebsiella variicola*	*Klebsiella*	*Klebsiella*

Table 2. Identity of the most closely related species (genus) for the isolates characterized here. Species (genus) identification was performed using average nucleotide identity (ANI), 16S rRNA and *nifH* sequence comparisons. [a]NA—not applicable, since these isolates do not encode *nifH* genes.

nifH genes in these four genomes. This could be due to false-positives in the original PCR analysis for the presence of *nifH* genes, or to changes in the composition of (possibly mixed) bacterial cultures during subsequent growth steps after the initial isolation on nitrogen-free media.

Most of the isolates characterized here belong to the genus *Klebsiella. K. variicola,* 14 of 22 isolates, is the most abundant species characterized here. This finding is consistent with previous studies showing that *Klebsiella* strains are capable of fixing nitrogen[12]; in fact, the canonical *nif* operons were defined in the *K. variicola* type strain 342 (originally classified as *Kelbsiella pneumoniae* strain 342) genome sequence[13].

K. variicola is also known to be an opportunistic pathogen that can cause disease in immunocompromised human hosts[14,15], which raises obvious safety concerns regarding its application to crops as part of a biofertilizer inoculum. We performed a comparative sequence analysis between the endophytic nitrogen-fixing *K. variicola* type strain 342, which is capable of infecting the mouse urinary tract and lung[13], and five of the isolates identified as *K. variicola* here. The *nif* cluster, which contains five functionally related *nif* operons involved in nitrogen fixation, is present in all of these genomes (Fig. 2). However, the four most critical pathogenicity islands implicated in the virulence of *K. variicola* 342 are all missing in the environmental *K. variicola* isolates characterized here (PAI 1–4 in Fig. 2a). The absence of pathogenicity islands in the genome of the endophytic nitrogen-fixer *K. michiganensis* Kd70 is associated with an inability to infect the urinary tract in mice[16]. Our results indicate that nitrogen-fixing *K. variicola* environmental isolates from Colombian sugarcane fields do not pose a health risk compared to clinical and environmental isolates that have previously been associated with pathogenicity. We explore this possibility in more detail in the following section on computational phenotyping.

The *nifH* genes from the *Klebsiella* isolates characterized here form two distinct phylogenetic clusters (Fig. 3). This finding is consistent with previous results showing multiple clades of *nifH* among *Klebsiella* genome sequences[17–19] and underscores the potential functional diversity, with respect to nitrogen fixation, for the sugarcane isolates characterized here.

Computational phenotyping. Computational phenotyping, also referred to as reverse genomics, was used to evaluate the potential of the bacterial isolates characterized here to serve as biofertilizers for Colombian sugarcane fields. For the purpose of this study, computational phenotyping entails the prediction of specific organismal phenotypes, or biochemical capacities, based on the analysis of functionally annotated genome sequences[20]. The goal of the computational phenotyping performed here was to identify isolates that show the highest predicted capacity for plant growth promotion while presenting the lowest risk to human populations. Accordingly, bacterial isolate genome sequences were screened for gene features that correspond to the desirable (positive) characteristics of (1) nitrogen fixation and (2) plant growth promotion and the disadvantageous (negative) characteristics of (3) virulence and (4) antimicrobial resistance. Genome sequences were scored and ranked according to the combined presence or absence of these four categories of gene features as described in the "Methods". To compute genome scores, the presence of nitrogenase and plant growth promoting genes

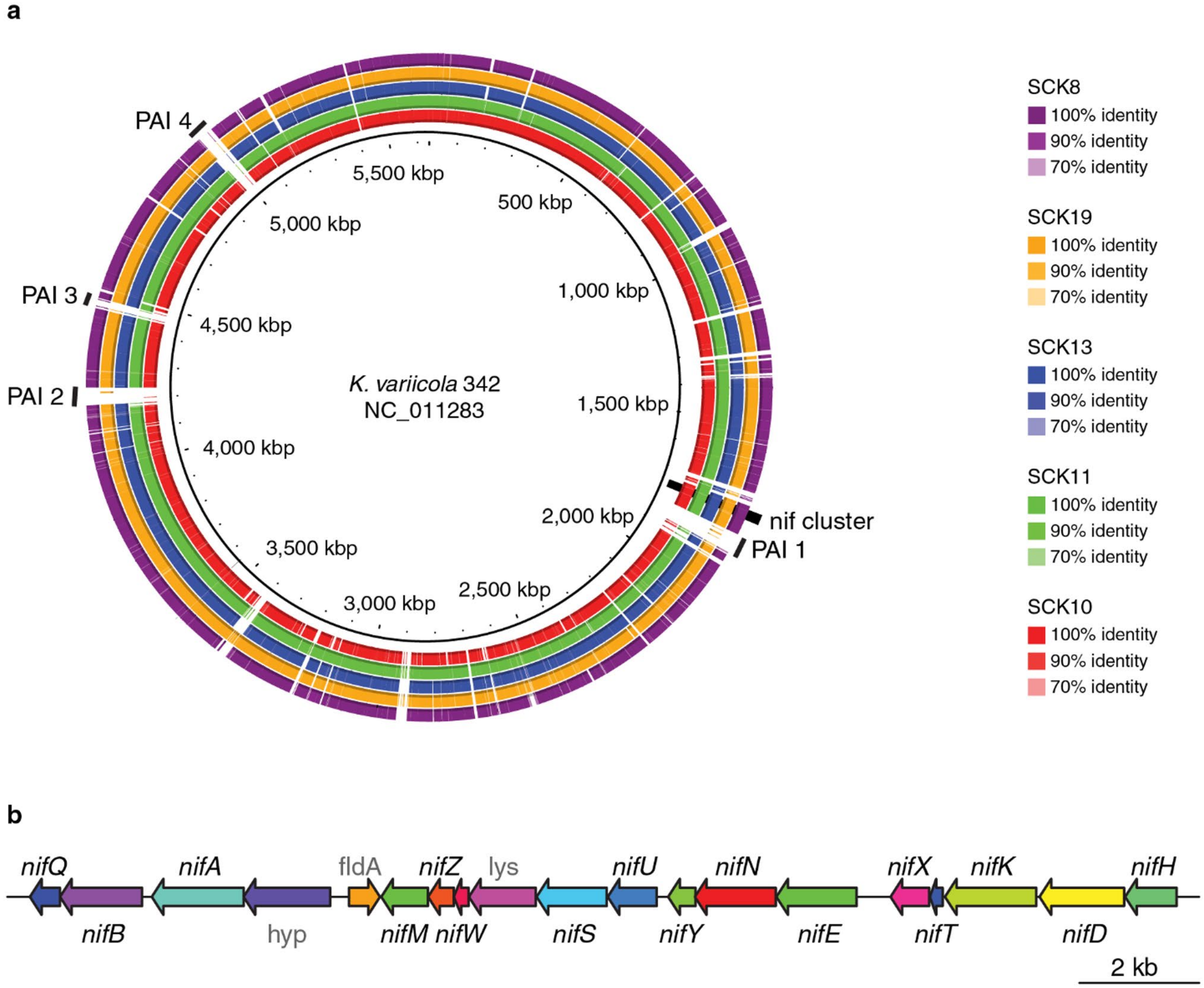

Figure 2. Comparison of the *K. variicola* type strain 342 to *K. variicola* sugarcane isolates characterized here. (**a**) BLAST ring plot showing synteny and sequence similarity between *K. variicola* 342 and five *K. variicola* sugarcane isolates. The *K. variicola* 342 genome sequence is shown as the inner ring, and syntenic regions of the five *K. variicola* sugarcane isolates are shown as rings with strain-specific color-coding according to the percent identity between regions of *K. variicola* 342 and the sugarcane isolates. The genomic locations of *nif* operon cluster along with four important pathogenicity islands (PAIs) are indicated. PAI1—type IV secretion and aminoglycoside resistance, PAI2 hemolysin and fimbria secretion, heme scavenging, PAI3—radical S-adenosyl-L-methionine (SAM) and antibiotic resistance pathways, PAI4—fosfomycin resistance and hemolysin production. (**b**) A scheme of the *nif* operon cluster present in both *K. variicola* 342 and the five *K. variicola* sugarcane isolates.

contribute positive values, whereas the presence of virulence factors and predicted antibiotic resistance yield negative values. Scores for each of the four specific phenotypic categories were normalized and combined to yield a single composite score for each bacterial isolate genome. The highest scoring isolates are predicted as best candidates to be included as part of a sugarcane biofertilizer inoculum. The predicted biochemical capacities of the highest scoring isolates were subsequently experimentally validated.

The results of the computational phenotyping analysis for the 22 bacterial isolate genome sequences characterized here are visualized as a heatmap in Fig. 4, and the presence/absence patterns for all of the genes analyzed here are shown in Supplementary Table S2. Isolates are ranked according to their composite genome scores, with the highest potential (10.87) for biofertilizers production shown at the top. Individual gene and phenotype scores are color coded for each genome, and the four functional-specific categories are shown separately. Individual gene results are shown for the nitrogenase (*nif*) genes, whereas genes are combined into functional sub-categories for the plant growth promoting and virulence factor genes. Predicted antimicrobial resistance phenotypes are shown for individual antibiotic classes.

The *nif* gene presence/absence profiles are highly similar for all but four of the bacterial isolates characterized here. These four isolates do not correspond to the *Klebsiella* genus, or closely related species, and do not encode any *nif* genes. These four isolates represent bacterial species that are commonly found in soil[21–25], but they are

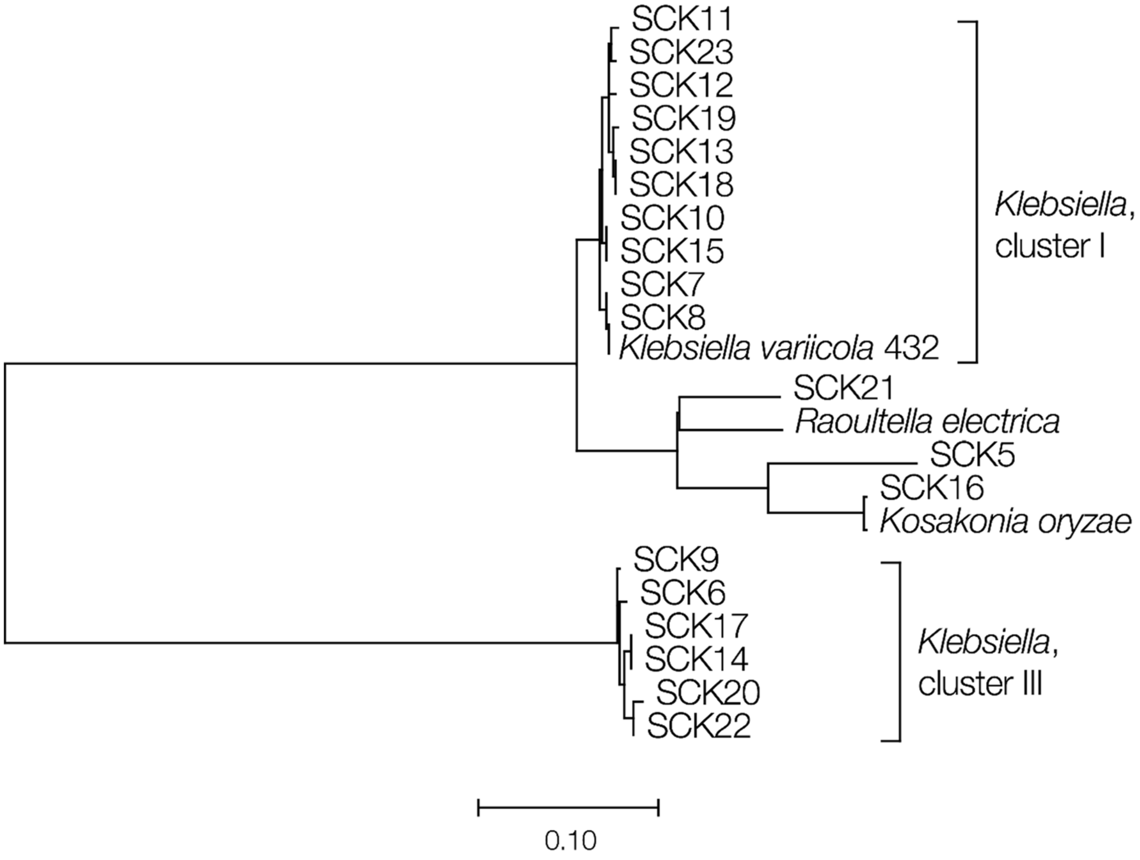

Figure 3. Phylogeny of the *nifH* genes for the bacterial isolates characterized here (SCK numbers). The phylogeny was reconstructed using pairwise nucleotide p-distances between *nifH* genes recovered from the isolate genome sequences using the neighbor-joining method. Horizontal branch lengths are scaled according the p-distances as shown.

not predicted to be viable biofertilizers. The *Kosakonia radicincitans* genome encodes the largest number of *nif* genes ($n = 17$) seen for any of the isolates characterized here. This is consistent with previous studies showing that isolates of this species can fix nitrogen[26]. The 14 *K.* variicola genomes characterized here all contain 16 out of 21 *nif* genes, including the core *nifD* and *nifK* genes, which encode the heterotetramer core of the nitrogenase enzyme, and the *nifH* gene, which encodes the dinitrogenase reductase subunit[27]. These genomes also all encode the nitrogenase master regulators *nifA* and *nifL*. The missing *nif* genes for the *K. variicola* isolates correspond to accessory structural and regulatory proteins that are not critical for nitrogen fixation. Accordingly, all of *K. variicola* isolate genomes are predicted to encode the capacity for nitrogen fixation, consistent with previous results[13,28]. The single *Raoultella electrica* isolate characterized here also contains the same 16 *nif* genes; *Raoultella* species have previously been isolated from sugarcane[29] and have also been demonstrated to fix nitrogen[30].

Initially, a total of 29 canonical bacterial plant growth promoting genes were mined from the literature, 25 of which were found to be present in at least one of the bacterial isolate genome sequences characterized here. These 25 plant growth promoting genes were organized into six distinct functional categories: phosphate solubilization, indolic acetic acid (IAA) production, siderophore production, 1-aminocyclopropane-1-carboxylate (ACC) deaminase, acetoin butanediol synthesis, and peroxidases (Supplementary Table S3). For the purposes of visualization (Fig. 4), each functional category is deemed to be present in an isolate genome sequence if all required genes for that function can be found, but the weighted scoring for these categories is based on individual gene counts as described in the "Methods". The *R. electrica* isolate shows the highest predicted capacity for plant growth promotion, with 5 of the 6 functional categories found to be fully present. The majority *K. variicola* isolates also show similar, but not identical, plant growth promoting gene presence/absence profiles, with 3 or 4 functional categories present. The capacity for siderophore production is predicted to vary among *K. variicola* isolates. The *K. radicincitans* genome also encodes 4 functional categories of plant growth promoting genes, but differs from the *K. variicola* isolates with respect to absence of phosphate solubilization genes and the presence of acetoin butanediol synthesis genes. Three of the four species found to lack *nif* genes also do not score present for any of the plant growth promoting gene categories, further underscoring their predicted lack of utility as biofertilizers.

Initially, a total of ~ 2500 virulence factor genes were mined from the Virulence Factor Database (VFDB)[31], 44 of which were found to be present in at least one of the bacterial isolate genome sequences characterized here. These 44 virulence factors were organized into six distinct functional categories related to virulence and toxicity: adherence, invasion, capsules, endotoxins, exotoxins, and siderophores. The weighted scores for these categories were computed based on individual gene presence/absence patterns, which were combined to yield the color

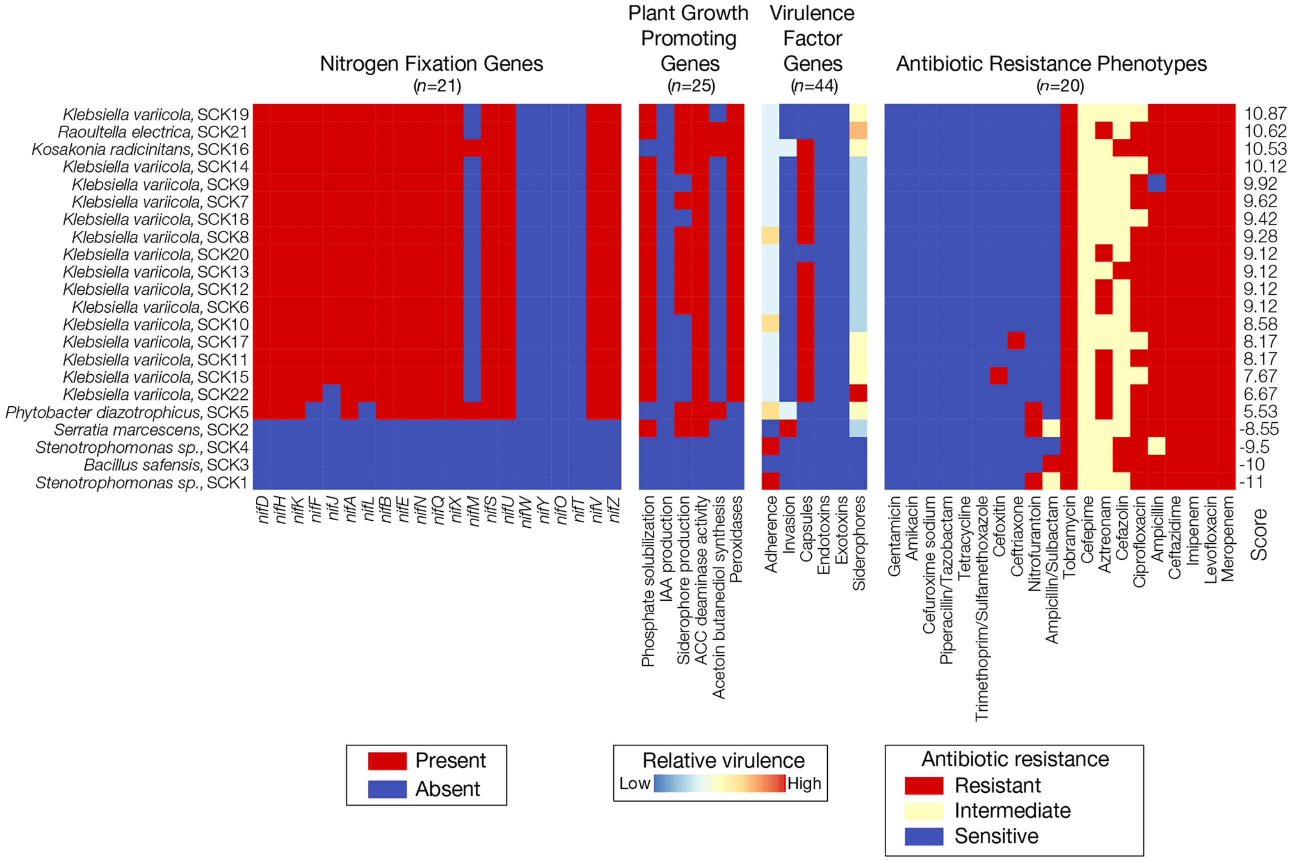

Figure 4. Computational phenotyping of the sugarcane bacterial isolates characterized here. The presence (red) and absence (blue) profiles for nitrogen fixation genes, plant growth promoting genes, and virulence factor genes are shown for the 22 bacterial isolates. Results are shown for all $n = 21$ nitrogen-fixing genes. Results for plant growth promoting genes ($n = 25$) and virulence factor genes ($n = 44$) are merged into six gene categories each. Predicted antibiotic resistance profiles are shown for $n = 20$ antibiotic classes. Detailed results for gene presence/absence and predicted antibiotic resistance profiles are shown in Supplementary Table S2. The results for all four phenotypic classes of interest were merged into a single priority score for each isolate (right side of plot), as described in the "Methods", and used to rank the isolates with respect to their potential as biofertilizers.

scheme shown in Fig. 4. Despite the fact that *K. pneumoniae* clinical isolates have previously been characterized as opportunistic pathogens, the *K. variicola* environmental isolates characterized here show uniformly low virulence scores. The virulence factor genes found among the *K. variicola* isolates correspond to adherence proteins, capsules, and siderophores. As shown in Fig. 2, these genomes lack coding capacity for important invasion and toxin proteins, including the Type IV secretion system, which can be found in clinical *K. pneumoniae* isolates. The *R. electrica* and *K. radicincitans* isolates, both of which show high scores for nitrogen fixation and plant growth promoting, have higher virulence scores than can be seen for the environmental *K. variicola* isolates characterized here. Whereas *Bacillus safensis* has the lowest virulence score for any of the isolates characterized here, the remaining three isolates that lack *nif* coding capacity have the highest virulence scores and encode well- known virulence factors, such as Type IV, hemolysin, and fimbria secretion systems. The results of a more detailed comparison of the predicted virulence for clinical isolates of *K. pneumoniae* and closely related species, compared to the environmental isolates, are reported in the following section of the manuscript.

The predicted antibiotic resistance phenotypes for the 20 classes of antimicrobial compounds for which predictions were made are fairly similar across the isolates characterized here. The majority of the *K. variicola* isolates, along with the relatively high scoring *R. electrica* and *K. radicincitans* isolates, show predicted susceptibility to 10 of the 20 classes of antimicrobial compounds, intermediate susceptibility for 2–4, and predicted resistance to 5–8. The highest level of predicted antibiotic resistance was seen for *Serratia marcescens*, with resistance predicted for 8 compounds and intermediate susceptibility predicted for 4.

Computational phenotyping scores for the four categories were normalized and combined into a final score, which is shown on the right of Fig. 4 and used to rank the isolates top-to-bottom with respect to their potential as biofertilizers. Most of the top positions are occupied by *K. variicola* isolates, with the exception of the second-ranked *R. electrica* and the third-ranked *K. radicincitans*. The results of a similar analysis of four additional plant associated *Klebsiella* genomes are shown in Supplementary Figure S3.

Virulence comparison. The results described in the previous section indicate that the most of the *K. variicola* strains isolated from Colombian sugarcane fields have the highest overall potential as biofertilizers, including a low predicted potential for virulence. Nevertheless, the fact that strains of *K. variicola* have previously been characterized as opportunistic pathogens[17,32] raises concerns when considering the use of *K. variicola* as part of a bioinoculum that will be applied to sugarcane fields. With this in mind, we performed a broader comparison of the predicted virulence profiles for the 22 environmental bacterial isolates characterized here compared to a collection of 28 clinical and one environmental isolate of *K. pneumoniae* and several other closely related species (See Supplementary Table S5 for isolate accession numbers). For this comparison, the same virulence factor scoring scheme described in the previous section was applied to all 51 genome sequences. The results of this comparison are shown in Fig. 5. Perhaps most importantly, there is a very clear distinction in the virulence score distribution, whereby all surveyed clinical strains show higher predicted virulence (from 4.45 to 2.11) than any of the environmental isolates characterized here (1.55 to 0.00). Furthermore, the three environmental isolates that show the highest predicted virulence correspond to species with low predicted capacity for both nitrogen fixation and plant growth promotion; as such, these isolates are not being considered as potential biofertilizers. The *K. variicola* environmental isolates, on the other hand, show uniformly low predicted virulence compared to clinical isolates of the same species. These results support, in principle, the use of the environmental *K. variicola* isolates characterized here as biofertilizers for Colombian sugarcane fields.

Experimental validation of prioritized isolates. The top six scoring isolates from the computational phenotyping were subjected to a series of cultivation-based phenotypic assays in order to validate their predicted biochemical activities: (1) acetylene reduction (a proxy for nitrogen fixation), (2) phosphate solubilization, (3) siderophore production, (4) gibberellic acid production, and (5) indole acetic acid production.

Nitrogen fixation activity, as measured by acetylene reduction to ethylene, was observed in all six isolates, three of which had higher levels in comparison to the positive control (Fig. 6A). The remaining three isolates showed higher levels of ethylene production compared to the negative control. All six of the isolates showed high levels of phosphate solubilization (Fig. 6B, C) and siderophore production (Fig. 6D, E) compared to the respective negative controls. All six isolates showed the ability to produce gibberellic acid (Fig. 6F), whereas none were able to produce indole acetic acid. The biochemical assay results are consistent with the computational phenotype predictions for these isolates.

Discussion

The overall goal of this study was to characterize nitrogen-fixing bacteria and their potential as biofertilizers. Strains cultivated from Colombian sugarcane fields were screened by *nifH* specific PCR, and both 16S rRNA gene sequence and whole genome sequence comparisons were used to identify the isolates' taxonomic origins. We found that most of the isolates characterized in this study belong to the family *Enterobacteriaceae*, with *Klebsiella* as the most abundant genus. 15 of the 22 nitrogen-fixing bacteria cultivated from Colombian sugarcane were identified as *Klebsiella* (Fig. 1 and Table 2), and 7 distinct isolates from other genera were also identified, including *Raoultella electrica* and *Kosakonia radicincitans*, which are also members of the family *Enterobacteriaceae*. These two species are closely related to *Klebsiella* and have been previously misclassified as *Klebsiella*[33]. In addition, we also isolated *Serratia marcescens*, *Phytobacter diazotrophicus, Stenotrophomonas spp.*, and *Bacillus safensis* from Colombian sugarcane fields, all of which have previously been found in soils and associated with several plants, including sugarcane, and are opportunistic pathogens[34–37].

Klebsiella are Gram-negative, facultative anaerobic bacteria that can be isolated from soils, plants, or water[38]. *Klebsiella* species have been isolated from a large variety of crops worldwide, such as sugarcane, rice, wheat, and maize[38–40]. *Klebsiella* species associated with plants have been shown to fix nitrogen and express other plant growth promoting traits[12,39]. Many of the bacteria previously isolated from sugarcane fields belong to the family *Enterobacteriaceae*, and *Klebsiella* species are abundant amongst the cultivable strains of *Enterobacteriaceae* obtained from sugarcane[41]. *Klebsiella* species present in sugarcane fields have been identified in several areas of the world. A survey of sugarcane in Guangxi, China found *Enterobacteriaceae*, especially *Klebsiella*, to be the most abundant plant-associated nitrogen-fixing bacteria[41]. The same group demonstrated that a nitrogen-fixing strain of *K. variicola* was able to colonize sugarcane and promote plant growth[39]. In Brazil, endophytic *Klebsiella* spp. have been isolated from commercial sugarcane, and their ability to produce plant growth promoting activity has been evaluated in vitro[42]. In Pakistan, the phenotypic diversity of plant growth promoting bacteria associated with sugarcane has been determined, with *Klebsiella* also appearing as one of the most abundant bacteria found[43].

In this study, we developed a computational phenotyping approach for the screening of potential plant growth promoting bacteria that can serve as biofertilizers. Computational phenotyping entails the implementation of a variety of bioinformatic and statistical methods to predict phenotypes of interest based on whole genome sequence analysis[44,45]. This approach has been used for a variety of applications in the biomedical sciences: prediction of clinically relevant phenotypes, study of infectious diseases, identification of opportunistic pathogenic bacteria in the human microbiome, and cancer treatment decisions[46,47]. To our knowledge, this study represents the first-time computational phenotyping has been used for agricultural applications. To implement computational phenotyping for the prioritization of potential biofertilizers, we developed a scoring scheme based on the genome content of four functional gene categories of interest: nitrogen-fixing genes, other plant growth promoting genes, virulence factor genes, and antimicrobial resistance genes.

The results of the computational phenotyping predictions, confirmed by laboratory experiment, supported the potential use of some of the bacterial strains isolated from Colombian sugarcane fields as biofertilizers with minimal human health risk. This is particularly true for the isolates that show the higher scores (5.53–10.87, Fig. 4), all of which encode the potential to fix nitrogen and promote plant growth but lack many of the important

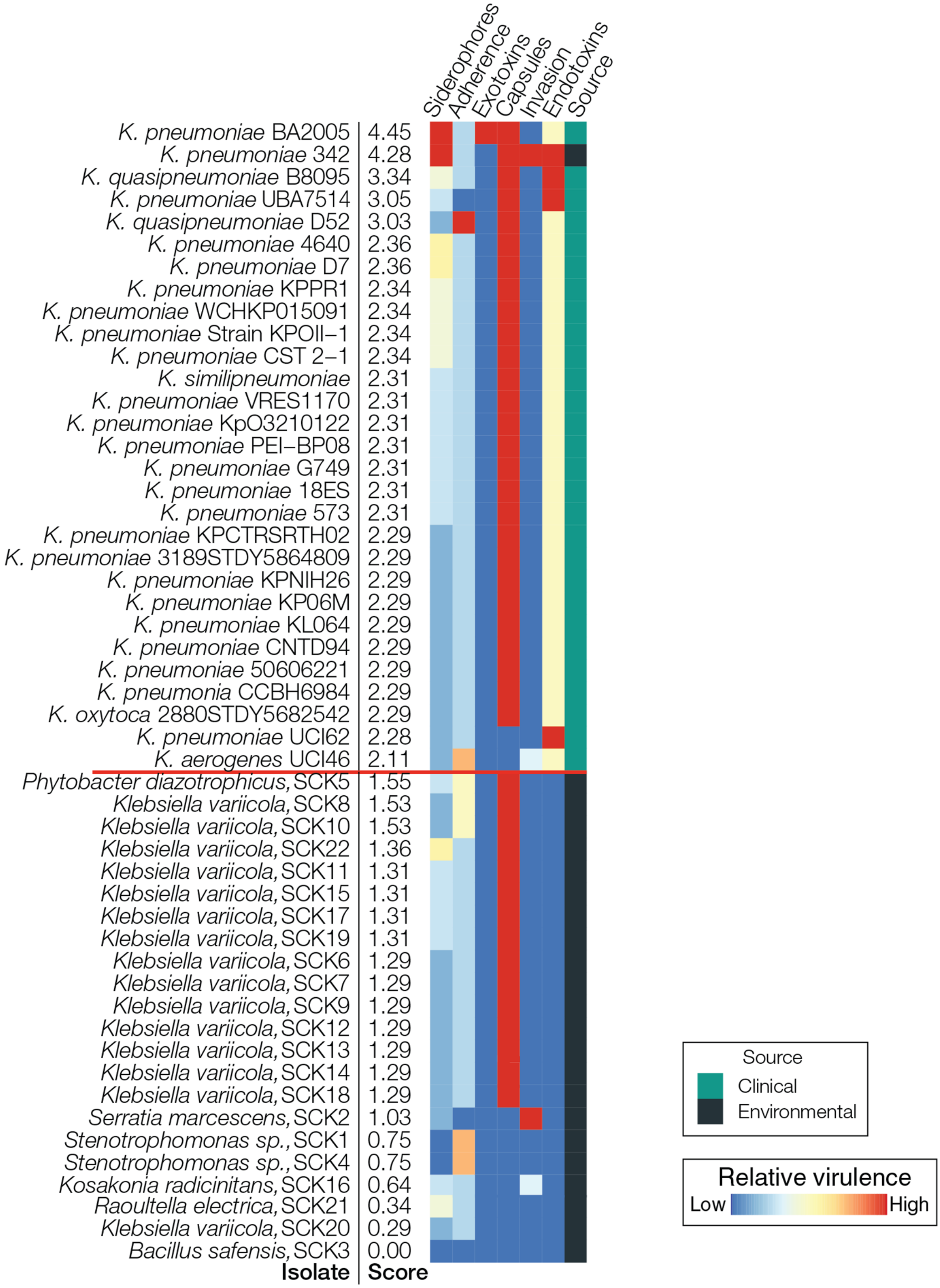

Figure 5. Comparison of predicted virulence profiles for clinical *K. pneumoniae* isolates compared to the environmental (sugarcane) bacterial isolates characterized here. As in Fig. 4, predicted virulence profiles for six classes of virulence factor genes are shown for each isolate. Isolate-specific virulence factor scores are shown for each isolate are based on the presence/absence profiles for the $n = 44$ virulence factor genes as described in the "Methods". Relative virulence levels are color-coded as shown in the key (note that the color coding here is slightly different than seen in Fig. 4). The virulence factor genes are used to rank the genomes from most (left) to least (right) virulent. Clinical versus environmental samples are shown to the left and right, respectively, of the red line, based on their virulence scores.

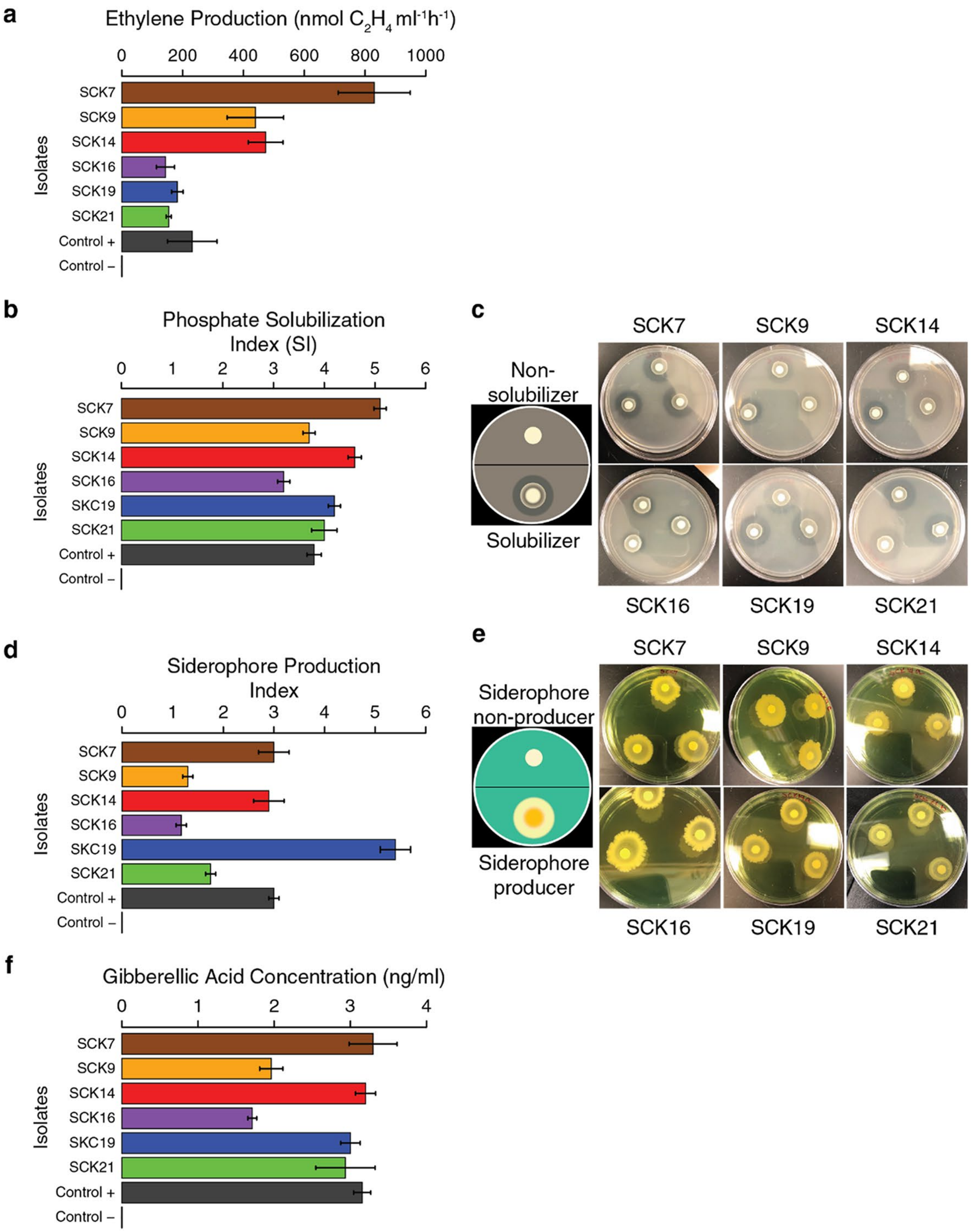

Figure 6. Experimental validation of prioritized biofertilizer isolates. The computationally predicted plant growth promoting phenotypes for the top six isolates were experimentally validated. All six strains were capable of acetylene reduction, i.e. ethylene production (*Azotobacter vinelandii* positive and *Escherichia coli* negative control) (**a**), phosphate solubilization (*Pseudomonas aeruginosa* ATCC 2785 positive and *E. coli* negative control) (**b**, **c**), siderophore production (*E. coli* ATCC 35,218 positive and *Bacillus safensis* negative control) (**d**, **e**), and gibberellic acid production (*A. vinelandii* positive and *Stenotrophomonas* sp negative control) (**f**). For panels (**a**), (**b**), (**d**), and (**f**), mean values are shown, with error bars representing + /− 1 standard deviation.

known virulence factors and antibiotic resistance genes that can be found in clinical isolates of the same species. In general, isolates SCK7, SCK14, and SCK19 appeared to possess more potent plant growth promoting properties compared to isolates SCK9, SCK16, and SCK21 (Fig. 4). Our computational phenotyping scheme also has valuable negative predictive value. We discovered isolates that show few or none of the beneficial traits that characterized biofertilizers; *Bacillus safensis* SCK3 and *Stenotrophomonas maltophilia* SCK1 had the lowest scores (− 10 and − 11 respectively). Finally, it is also worth reiterating that the computationally predicted biochemical activities related to plant growth promotion were all validated by experimental results (Fig. 6).

A number of the nitrogen-fixing bacterial species isolated from sugarcane can also be opportunistic pathogens, which are microorganisms that usually do not cause disease in a healthy host; instead, they colonize and infect the immunocompromised host[48,49]. Although *Klebsiella spp*. exist in the environment and show plant growth promoting potential, they have also been associated with nosocomial diseases in humans[17]. *Klebsiella* spp. isolates causing hospital-acquired infections, primarily in immunocompromised persons, include *Klebsiella pneumoniae*, *Klebsiella oxytoca*, and *Klebsiella granulomatis*[50].

The potential for virulence, along with the presence of antimicrobial resistance genes, is an obvious concern when proposing to use *Klebsiella* spp. as biofertilizers. Importantly, we found that the environmental *Klebsiella* isolates did not contain pathogenicity islands associated with many virulence factor genes usually found in clinical isolates of *Klebsiella* spp. (Fig. 2). These results are consistent with a recent study using whole genome sequences analysis[16], which found that the *Klebsiella michiganensis* Kd70 isolated from the intestine of larvae of *Diatraea saccharalis* contains multiple genes associated with plant growth promotion and root colonization, but lacks pathogenicity islands in its genome. We also performed a broader comparison of the presence of virulence factors in the environmental isolates characterized here versus genomes of *Klebsiella* clinical isolates associated with opportunistic infections in humans along with a number of other environmental isolates that have available genome sequences (Fig. 5). The virulence factor profiles for all of the environmental isolates were clearly distinct from the clinical strains, which show uniformly higher virulence profile scores, underscoring the relative safety of *Klebsiella* environmental isolates for use as biofertilizers.

The results obtained from the computational phenotyping approach developed in this study serve as a proof of principle in support of genomic guided approaches to sustainable agriculture. In particular, computational phenotyping can serve to substantially narrow the search space for potential plant growth promoting bacterial isolates, which can be further interrogated via experimental methods. Computational phenotyping can be used to simultaneously identify beneficial properties of plant associated bacterial isolates while avoiding potentially negative characteristics. In principle, this approach can be applied to a broad range of potential plant growth promoting isolates, or even assembled metagenomes, from managed agricultural ecosystems.

We can also envision a number of other potential applications for computational phenotyping of microbial genomes. The computational phenotyping methodology developed here has broad potential including diverse applications in agriculture, plant and animal breeding, food safety, water quality microbiology along with other industrial microbiology applications such as bioenergy, quality control/quality assurance, and fermentation microbiology as well as human health applications such as pathogen antibiotic resistance, virulence predictions, and microbiome characterization. For instance, computational phenotyping could be useful in food safety related to vegetable crop production. Vegetables such as lettuce, spinach, and carrots are usually consumed raw, which increases the potential for bacterial infections or human disease outbreaks[48,51]. Vegetable plants and other crops harbor diverse bacterial communities, and the dominant families in these communities vary according to different variables, such as soil type and host genotype. The microbiome of the sugar cane crops studied here is dominated by the family *Enterobacteriaceae*, Gram-negative bacteria that include a huge diversity of plant growth promoting bacteria and enteric pathogens[52–56]. However, in *Arabidopsis thaliana* and other plants, gamma-Proteobacteria other than *Enterobacteriaceae* appear to dominate the plant-associated microbiome[57].

Increasing antibiotic resistance, generated by the abuse of antibiotics in agriculture as well as medicine, is another major threat to human health[58], and the food supply chain creates a direct connection between the environmental habitat of bacteria and human consumers[59]. Our computational phenotyping approach could provide for an additional food safety solution, which could be used to prevent the spread of antibiotic resistant pathogens present in the food chain.

Conclusions

A genome-enabled approach was developed for the prioritization of native bacterial isolates with the potential to serve as biofertilizers for sugarcane fields in Colombia's Cauca Valley. The approach is based on computational phenotyping, which entails predictions related to traits of interest based on bioinformatic analysis of whole genome sequences. Bioinformatic predictions were validated through investigation of plant growth promoting traits with experimental assays in the laboratory, thereby demonstrating the utility of computational phenotyping for assessing the benefits and risks posed by bacterial isolates that can be used as biofertilizers. The quantitative approach to computational phenotyping developed here for the discovery of potential biofertilizers has broad potential applications for environmental and industrial microbiology, including potential use in food safety, water quality, and antibiotic resistance studies.

Methods

Sampling and cultivation of putative nitrogen-fixing bacteria from sugarcane. INCAUCA is a Colombian sugarcane company located in the Cauca River Valley in the southwest region of the country between the western and central Andes mountain ranges (http://www.incauca.com/). Samples of leaves, rhizosphere soil, stem, and roots were collected from the sugarcane fields designated as 32 T and 37 T of the INCAUCA San Fernando farm located in the Cauca Valley (3° 16′ 30.0″ N 76° 21′ 00.0″ W). Samples were collected with the permission of the United States Department of Agriculture and the Colombian Ministry of Environment and Sustainable Development, and all experiments were performed in compliance with institutional, national, and international guidelines. A high-throughput enrichment approach was developed to enable the cultivation of multiple strains of putative nitrogen-fixing bacteria from sugarcane field samples; details of this approach can be found in the Supplementary Material (Supplementary Methods and Supplementary Figure S1).

A total of 22 distinct *nifH* PCR + isolates that passed the initial cultivation and screening steps were grown in LB medium (Difco) at 37 °C for subsequent genomic DNA extraction. The E.Z.N.A. bacterial DNA kit (Omega Bio-Tek) was used for genomic DNA extraction, and paired-end fragment libraries (~ 1,000 bp) were constructed using the Nextera XT DNA library preparation kit (Illumina).

Genome sequencing, assembly, and annotation. Isolate genomic DNA libraries were sequenced on the Illumina MiSeq platform using V3 chemistry, yielding approximately 400,000 paired-end 300 bp sequence reads per sample. A list of all genome sequence analysis programs that were used for this study is provided Supplementary Table S4. Sequence read quality control and trimming were performed using the programs FastQC version0.11.5[60] and Trimmomatic (v.0.35)[61]. De novo sequence assembly was performed using the program SPAdes (v.3.6)[62]. Assembled genome sequences were annotated using the Rapid Annotations using Subsystems Technology (RAST) Web server[63,64] and NCBI Prokaryotic Genome Annotation Pipeline (PGAP)[65]. The 15 *Klebsiella* isolates characterized in this way were briefly described in a Genome Announcement[66], and the analysis here includes 7 additional non-*Klebesiella* isolates.

Comparative genomic analysis. Average Nucleotide Identity (ANI) was employed to assign the taxonomy of the bacterial isolates characterized here[67,68]. Taxonomic assignment was also conducted by targeting small subunit ribosomal RNA (SSU rRNA) gene sequences. Nitrogenase enzyme encoding *nifH* gene sequences were extracted from isolate genome sequences, clustered, and taxonomically assigned using the TaxaDiva (v.0.11.3) method developed by our group[11]. Whole genome sequence comparisons between bacterial isolates characterized here and the *K. variicola* type strain 342 were performed using BLAST + (v.2.2.28)[69] and visualized with the program CGView (v.1.0)[70]. Details of the methods used comparative genomic analysis can be found in the Supplementary Methods section.

Computational phenotyping. Computational phenotyping was performed by searching the bacterial isolate genome sequences characterized here for the presence/absence of genes or features related to four functional classes of interest, with respect to their potential as biofertilizers: (1) nitrogen fixation (NF), (2) plant growth promotion (PGP), (3) virulence factors, and (4) antimicrobial resistance (AMR). Gene panels were manually curated by searching the literature (NCBI PubMed) for genes implicated in nitrogen fixation and plant growth promotion. The Virulence Factors Database (VFDB) was used to curate the virulence factor gene panel[31]. AMR levels were quantified using the PATRIC3/mic prediction tool[71]. A composite score was developed to characterize each bacterial isolate genome sequence with respect to the presence/absence of genes from the NF, PGP, and VF gene panels along with the predicted AMR levels. Details on the gene panels, AMR level, and the composite scoring system can be found in the Supplementary Methods.

Experimental validation. Predictions made by computational phenotyping were validated using five distinct experimental assays: (1) Acetylene reduction assay for nitrogen fixation activity, (2) Phosphate solubilization assay, (3) Siderophore production assay, (4) Gibberellic acid production assay, and (5) Indole acetic acid production assay. Details of each experimental assay can be found in the Supplementary Methods.

Data availability

The datasets supporting the conclusions of this article are included within the article and its additional files. Sequencing and assembly data are available in the NCBI BioProject database under the accession PRJNA418312.

Received: 28 July 2020; Accepted: 7 April 2021
Published online: 28 April 2021

References

1. Fess, T. L., Kotcon, J. B. & Benedito, V. A. Crop breeding for low input agriculture: a sustainable response to feed a growing world population. *Sustainability-Basel* **3**, 1742–1772 (2011).
2. Bargaz, A., Lyamlouli, K., Chtouki, M., Zeroual, Y. & Dhiba, D. Soil microbial resources for improving fertilizers efficiency in an integrated plant nutrient management system. *Front. Microbiol.* **9**, 1606. https://doi.org/10.3389/fmicb.2018.01606 (2018).
3. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. U S A* **108**, 20260–20264. https://doi.org/10.1073/pnas.1116437108 (2011).
4. Stewart, W. M., Dibb, D. W., Johnston, A. E. & Smyth, T. J. The contribution of commercial fertilizer nutrients to food production. *Agron. J.* **97**, 1–6 (2005).
5. Savci, S. Investigation of effect of chemical fertilizers on environment. *Int. Conf. Environ. Sci. Dev.* **1**, 287–292 (2012).
6. Bhardwaj, D., Ansari, M. W., Sahoo, R. K. & Tuteja, N. Biofertilizers function as key player in sustainable agriculture by improving soil fertility, plant tolerance and crop productivity. *Microb. Cell Fact.* https://doi.org/10.1186/1475-2859-13-66 (2014).
7. Cherubin, M. R. *et al.* Soil quality indexing strategies for evaluating sugarcane expansion in Brazil. *PLoS ONE* **11**, e0150860. https://doi.org/10.1371/journal.pone.0150860 (2016).
8. Selman-Housein, G. *et al.* Towards the improvement of sugarcane bagasse as raw material for the production of paper pulp and animal feed. *Dev. Plant Genet.* **5**, 189–193 (2000).
9. Dong, M. *et al.* Diversity of the bacterial microbiome in the roots of four Saccharum species: *S. spontaneum*, *S. robustum*, *S. barberi*, and *S. officinarum*. *Front. Microbiol.* **9**, 267 (2018).
10. Li, H. B. *et al.* Genetic diversity of nitrogen-fixing and plant growth promoting *Pseudomonas* species isolated from sugarcane rhizosphere. *Front. Microbiol.* **8**, 1268 (2017).
11. Gaby, J. C. *et al.* Diazotroph community characterization via a high-throughput *nifH* amplicon sequencing and analysis pipeline. *Appl. Environ. Microbiol.* https://doi.org/10.1128/AEM.01512-17 (2018).

12. Lin, L. *et al.* Complete genome sequence of endophytic nitrogen-fixing *Klebsiella variicola* strain DX120E. *Stand. Genom. Sci.* **10**, 22. https://doi.org/10.1186/s40793-015-0004-2 (2015).
13. Fouts, D. E. *et al.* Complete genome sequence of the N_2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. *PLoS Genet.* **4**, e1000141. https://doi.org/10.1371/journal.pgen.1000141 (2008).
14. Rodriguez-Medina, N., Barrios-Camacho, H., Duran-Bedolla, J. & Garza-Ramos, U. *Klebsiella variicola*: an emerging pathogen in humans. *Emerg. Microbes Infect.* **8**, 973–988. https://doi.org/10.1080/22221751.2019.1634981 (2019).
15. Berry, G. J., Loeffelholz, M. J. & Williams-Bouyer, N. An investigation into laboratory misidentification of a bloodstream *Klebsiella variicola* infection. *J. Clin. Microbiol.* **53**, 2793–2794. https://doi.org/10.1128/JCM.00841-15 (2015).
16. Dantur, K. I. *et al.* The endophytic strain *Klebsiella michiganensis* Kd70 lacks pathogenic island-like regions in its genome and is incapable of infecting the urinary tract in mice. *Front. Microbiol.* **9**, 1548. https://doi.org/10.3389/fmicb.2018.01548 (2018).
17. Rosenblueth, M., Martinez, L., Silva, J. & Martinez-Romero, E. *Klebsiella variicola*, a novel species with clinical and plant-associated isolates. *Syst. Appl. Microbiol.* **27**, 27–35. https://doi.org/10.1078/0723-2020-00261 (2004).
18. Raymond, J., Siefert, J. L., Staples, C. R. & Blankenship, R. E. The natural history of nitrogen fixation. *Mol. Biol. Evol.* **21**, 541–554. https://doi.org/10.1093/molbev/msh047 (2004).
19. Zehr, J. P., Jenkins, B. D., Short, S. M. & Steward, G. F. Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ. Microbiol.* **5**, 539–554 (2003).
20. Weimann, A. *et al.* From genomes to phenotypes: Traitar, the microbial trait analyzer. *mSystems* **1**, e00101-00116. https://doi.org/10.1128/mSystems.00101-16 (2016).
21. Deredjian, A. *et al.* Occurrence of *Stenotrophomonas maltophilia* in agricultural soils and antibiotic resistance properties. *Res. Microbiol.* **167**, 313–324. https://doi.org/10.1016/j.resmic.2016.01.001 (2016).
22. Caulier, S. *et al.* Versatile antagonistic activities of soil-borne *Bacillus* spp. and *Pseudomonas* spp. against *Phytophthora infestans* and other potato pathogens. *Front. Microbiol.* **9**, 143. https://doi.org/10.3389/fmicb.2018.00143 (2018).
23. Badran, S. *et al.* Complete genome sequence of the *Bacillus pumilus* phage Leo2. *Genome Announc.* https://doi.org/10.1128/genomeA.00066-18 (2018).
24. Pavan, M. E. *et al.* Phylogenetic relationships of the genus *Kluyvera*: transfer of *Enterobacter intermedius* Izard et al. 1980 to the genus *Kluyvera* as *Kluyvera intermedia* comb. nov. and reclassification of *Kluyvera cochleae* as a later synonym of *K. intermedia*. *Int. J. Syst. Evol. Microbiol.* **55**, 437–442. https://doi.org/10.1099/ijs.0.63071-0 (2005).
25. Zhang, G. X. *et al.* Diverse endophytic nitrogen-fixing bacteria isolated from wild rice *Oryza rufipogon* and description of *Phytobacter diazotrophicus* gen. nov. sp. nov. *Arch. Microbiol.* **189**, 431–439. https://doi.org/10.1007/s00203-007-0333-7 (2008).
26. Berger, B., Wiesner, M., Brock, A. K., Schreiner, M. & Ruppel, S. K. *radicincitans*, a beneficial bacteria that promotes radish growth under field conditions. *Agron. Sustain. Dev.* **35**, 1521–1528. https://doi.org/10.1007/s13593-015-0324-z (2015).
27. Stacey, G., Burris, R. H. & Evans, H. J. *Biological Nitrogen Fixation* (Chapman and Hall, 1992).
28. Scott, K. F., Rolfe, B. G. & Shine, J. Biological nitrogen fixation: primary structure of the *Klebsiella pneumoniae nifH* and *nifD* genes. *J. Mol. Appl. Genet.* **1**, 71–81 (1981).
29. Luo, T. *et al. Raoultella* sp. strain L03 fixes N_2 in association with micropropagated sugarcane plants. *J. Basic Microbiol.* **56**, 934–940. https://doi.org/10.1002/jobm.201500738 (2016).
30. Schicklberger, M., Shapiro, N., Loque, D., Woyke, T. & Chakraborty, R. Draft genome sequence of *Raoultella terrigena* R1Gly, a diazotrophic endophyte. *Genome Announc.* https://doi.org/10.1128/genomeA.00607-15 (2015).
31. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. Hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res.* **44**, D694-697 (2016).
32. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl. Acad. Sci. USA* **112**, E3574–E3581. https://doi.org/10.1073/pnas.1501049112 (2015).
33. Drancourt, M., Bollet, C., Carta, A. & Rousselier, P. Phylogenetic analyses of Klebsiella species delineate Klebsiella and Raoultella gen. nov., with description of *Raoultella ornithinolytica* comb. Nov., *Raoultella terrigena* comb. nov. and *Raoultella planticola* comb. nov. *Int. J. Syst. Evol. Microbiol.* **51**, 925–932. https://doi.org/10.1099/00207713-51-3-925 (2001).
34. Denton, M. & Kerr, K. G. Microbiological and clinical aspects of infection associated with *Stenotrophomonas maltophilia*. *Clin. Microbiol. Rev.* **11**, 57–80 (1998).
35. Downing, K. J., Leslie, G. & Thomson, J. A. Biocontrol of the sugarcane borer *Eldana saccharina* by expression of the *Bacillus thuringiensis* cry1Ac7 and *Serratia marcescens* chiA genes in sugarcane-associated bacteria. *Appl. Environ. Microbiol.* **66**, 2804–2810. https://doi.org/10.1128/aem.66.7.2804-2810.2000 (2000).
36. Ribeiro, V. B. *et al.* Detection of bla(GES-5) in carbapenem-resistant Kluyvera intermedia isolates recovered from the hospital environment. *Antimicrob. Agents Chemother.* **58**, 622–623. https://doi.org/10.1128/AAC.02271-13 (2014).
37. Juhnke, M. E. & des Jardin, E. ,. Selective medium for isolation of *Xanthomonas maltophilia* from soil and rhizosphere environments. *Appl. Environ. Microbiol.* **55**, 747–750 (1989).
38. Bagley, S. T. Habitat association of *Klebsiella* species. *Infect. Control* **6**, 52–58 (1985).
39. Wei, C. Y. *et al.* Endophytic nitrogen-fixing *Klebsiella variicola* strain DX120E promotes sugarcane growth. *Biol. Fertil. Soils* **50**, 657–666 (2014).
40. Ji, S. H., Gururani, M. A. & Chun, S. C. Isolation and characterization of plant growth promoting endophytic diazotrophic bacteria from Korean rice cultivars. *Microbiol. Res.* **169**, 83–98. https://doi.org/10.1016/j.micres.2013.06.003 (2014).
41. Lin, L. *et al.* Plant growth-promoting nitrogen-fixing enterobacteria are in association with sugarcane plants growing in Guangxi, China . *Microbes Environ.* **27**, 391–398 (2012).
42. Beneduzi, A. *et al.* Diversity and plant growth promoting evaluation abilities of bacteria isolated from sugarcane cultivated in the South of Brazil. *Appl. Soil Ecol.* **63**, 94–104 (2013).
43. Mehnaz, S., Baig, D. N. & Lazarovits, G. Genetic and phenotypic diversity of plant growth promoting rhizobacteria isolated from sugarcane plants growing in Pakistan. *J. Microbiol. Biotechnol.* **20**, 1614–1623 (2010).
44. Richesson, R. L., Sun, J. M., Pathak, J., Kho, A. N. & Denny, J. C. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif. Intell. Med.* **71**, 57–61 (2016).
45. Drouin, A. *et al.* Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* https://doi.org/10.1186/s12864-016-2889-6 (2016).
46. Berger, A. H. *et al.* High-throughput phenotyping of lung cancer somatic mutations. *Cancer Res.* **30**, 214–228 (2016).
47. Bone, W. P. *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* **18**, 608–617 (2016).
48. Berg, G., Erlacher, A., Smalla, K. & Krause, R. Vegetable microbiomes: is there a connection among opportunistic infections, human health and our "gut feeling"?. *Microb. Biotechnol.* **7**, 487–495. https://doi.org/10.1111/1751-7915.12159 (2014).
49. Fishman, J. A. Opportunistic infections–coming to the limits of immunosuppression?. *Cold Spring Harb. Perspect. Med.* **3**, a015669. https://doi.org/10.1101/cshperspect.a015669 (2013).
50. Podschun, R. & Ullmann, U. Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin. Microbiol. Rev.* **11**, 589–603 (1998).
51. Osterblad, M., Pensala, O., Peterzens, M., Heleniusc, H. & Huovinen, P. Antimicrobial susceptibility of Enterobacteriaceae isolated from vegetables. *J. Antimicrob. Chemother.* **43**, 503–509. https://doi.org/10.1093/jac/43.4.503 (1999).

52. Yeoh, Y. K. *et al.* The core root microbiome of sugarcanes cultivated under varying nitrogen fertilizer application. *Environ. Microbiol.* **18**, 1338–1351. https://doi.org/10.1111/1462-2920.12925 (2016).
53. Magnani, G. S. *et al.* Diversity of endophytic bacteria in Brazilian sugarcane. *Genet. Mol. Res.* **9**, 250–258. https://doi.org/10.4238/vol9-1gmr703 (2010).
54. Dos-Santos, C. M. *et al.* A culture-independent approach to enrich endophytic bacterial cells from sugarcane stems for community characterization. *Microb. Ecol.* **74**, 453–465. https://doi.org/10.1007/s00248-017-0941-y (2017).
55. de Souza, R. S. *et al.* Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. *Sci. Rep.* **6**, 28774. https://doi.org/10.1038/srep28774 (2016).
56. de Santi Ferrara, F. I., Oliveira, Z. M., Gonzales, H. H. S., Floh, E. I. S. & Barbosa, H. R. Endophytic and rhizospheric enterobacteria isolated from sugar cane have different potentials for producing plant growth-promoting substances. *Plant Soil* **353**, 409–417 (2012).
57. Lundberg, D. S. *et al.* Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**, 86–90. https://doi.org/10.1038/nature11237 (2012).
58. Canica, M., Manageiro, V., Abriouel, H., Moran-Gilad, J. & Franz, C. M. A. P. Antibiotic resistance in foodborne bacteria. *Trends Food Sci. Technol.* **84**, 41–44 (2019).
59. Bengtsson-Palme, J. Antibiotic resistance in the food supply chain: where can sequencing and metagenomics aid risk assessment?. *Curr. Opin. Food Sci.* **14**, 66–71 (2017).
60. Andrews, S. *FastQC a quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
62. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
63. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genom.* **9**, 75 (2008).
64. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2013).
65. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624. https://doi.org/10.1093/nar/gkw569 (2016).
66. Medina-Cordoba, L. K. *et al.* Genome sequences of 15 *Klebsiella* sp. isolates from sugarcane fields in Colombia's Cauca Valley. *Genome Announc.* https://doi.org/10.1128/genomeA.00104-18 (2018).
67. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 2567–2572. https://doi.org/10.1073/pnas.0409727102 (2005).
68. Goris, J. *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91. https://doi.org/10.1099/ijs.0.64483-0 (2007).
69. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421. https://doi.org/10.1186/1471-2105-10-421 (2009).
70. Grant, J. R., Arantes, A. S. & Stothard, P. Comparing thousands of circular genomes using the CGView comparison tool. *BMC Genom.* **13**, 202. https://doi.org/10.1186/1471-2164-13-202 (2012).
71. Nguyen, M. *et al.* Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.* **8**, 421. https://doi.org/10.1038/s41598-017-18972-w (2018).

Acknowledgements

We thank members of the INCAUCA Laboratory of Microorganismal Production for their support with the isolation of sugarcane-associated bacteria in Colombia. We thank the Kostka laboratory technicians Patrick Steck and Michael Blejwas for their support with laboratory analysis of sugarcane-associated bacteria at Georgia Tech.

Author contributions

L.K.M., A.T.C., and L.R. analyzed, interpreted, and visualized the sequencing data, computational phenotyping, and experimental assay results. L.K.M., A.T.C., L.W.M., and J.E.K. designed the phenotypic screening experimental assays. L.K.M. performed all experimental assays. L.K.M., L.C.V., J.C.G., J.E.K. collected and isolated the bacterial isolates used in this study. A.V., J.E.K., and I.K.J. designed and supervised this study. J.E.K. and I.K.J. provided funding. L.K.M., A.T.C., I.K.J., J.E.K. wrote the manuscript. All authors read and approved the final manuscript.

Funding

Partial financial support was provided by the National Science Foundation (DEB-1754756).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-88380-8.

Correspondence and requests for materials should be addressed to J.E.K. or I.K.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.