Programmed strain tagging and gene disruption throughout a diverse bacterial genus

Ezgi Mehmetoğlu Boz[1], Hugo R. Barajas*[1], Tsung-Ta Yu*[1], Mounika Thigulla*[1,2], Naila Nazir[1], Sarfaraz Hussain#[1], Lorena Carot Hernández#[1,3], and Derek Severi Lundberg ϕ[1].

[1]Department of Plant Biology, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden

[2]Crop Biotechnics, KU Leuven, 3001 Leuven, Belgium

*,# = equal contribution

ϕ Corresponding author derek.lundberg@slu.se (D.S.L.)

Abstract

Transposons are a convenient vehicle to insert DNA into a bacterial genome, but widely-used transposons such as Tn7 are not able to hit custom targets. Recently, CRISPR RNA-guided integrases have been shown to direct the insertion of a mini transposon to a chosen site using a short guide sequence. We adapted this system for the widespread plant-associated genus *Sphingomonas*, revealing flexibilities and limitations of the tool. We uniquely tagged five genetically diverse strains, both in neutral sites and in sites with phenotypic consequences, and we demonstrate the utility of the tags for quantitative strain tracking in complex bacterial populations. Although we initially focused on the attTn7 site recognized by Tn7 as an insertion site with minimal fitness effects, we discovered via a genus-wide search that this site disrupts potentially important genes in some strains. Therefore we validated an improved site for benign integration in *Sphingomonas* and used a construct with guides targeting this site to transform a heterogeneous *Sphingomonas* population, bypassing prior strain isolation. Using a novel rapid and economical transposon mapping method, we were able to identify correctly-tagged primary transformant colonies with novel genetic content, thus demonstrating a short cut towards the establishment of diverse tagged synthetic communities for the experimental study of bacterial natural variation.

Introduction

To study a particular organism in its environment, it is important not to mistake that organism for a similar "look alike". Yet, distinguishing one organism from another can be challenging, especially for organisms like bacteria that often grow in complex, heterogenous communities. While a particular organism can sometimes be identified by its unique genotype using DNA sequencing, innate genetic differences are not helpful when an organism is genetically very similar or even identical to others in the population. Consequently, experiments involving synthetic bacterial communities must be carefully controlled to avoid external contamination, and the requirement that community members are distinguishable from each other limits community size and environmental complexity [1].

As an alternative to detecting subtle innate features, a common approach is to unambiguously modify an organism with a unique genetic element such as a fluorescent protein, an antibiotic resistance marker, or a DNA sequence that serves as a barcode. This approach of tagging bacteria, particularly through the use of antibiotic resistance and fluorescence, has been integral to the study of host-microbe interactions [2–4], and has allowed detailed characterization of individual genotypes in complex environments. However, it is increasingly of interest to track many genotypes simultaneously, and while antibiotic resistances and fluorescence proteins are limited in variety [5], DNA barcodes can distinguish many more individuals.

While genetic modification of some organisms like *E.coli* has become very efficient, inserting DNA barcodes into many non-model organisms remains a rate-limiting step. Bacterial transposons are an attractive high-throughput method to introduce DNA because they are functional in diverse bacteria, can carry a large payloads involving thousands of base pairs, and can integrate into the existing genome, making them less transient than broad host range plasmids which can be lost [6–8]. However, most transposons have insertion preferences that are not programmable, limiting their versatility.

The lack of programmability is especially problematic because tags should ideally have minimal impact on bacterial behavior and physiology, and transposon integration into the wrong location risks mutating important functions. Recent efforts have made progress developing tagging systems that can in principle introduce unique tags into large and diverse collections of bacteria [6,7,9], and these efforts almost always employ the Tn7 transposon because it recognizes a conserved integration site called attTn7, guiding insertions just downstream of the Glutamine--fructose-6-phosphate aminotransferase (*glmS*) gene in diverse Gram-negative bacteria. In most cases, Tn7 integrations are not associated with fitness penalties [7,10].

Like Tn7, CRISPR RNA-guided integrases, in particular the type I-F *Vibrio cholerae* CAST (VchCAST) Tn6677-like elements [8,11], are functional in diverse bacteria, and they have the major advantage that the integration site can be programmed using a guide RNA comprising a

5′ CN PAM site immediately followed by a 32 bp sequence, with the transposon usually integrating 50 bp downstream of the guide sequence. In this work, we adapted a single vector CRISPR RNA-guided integrase system [8] for broad use in diverse strains of the widespread plant-associated bacterium *Sphingomonas*, Gram-negative Alphaproteobacteria that grow in highly heterogeneous communities and researched for beneficial features including plant growth promotion, xenobiotic degradation [12–14]. We created a unique, minimalist tag, and tagged five diverse *Sphingomonas* strains in multiple genomic locations which we characterized in detail to reveal flexibilities and limitations of the system, and we validated the tags for quantitative strain tracking in complex environments. Importantly, we discovered that the popular attTn7 site is unlikely to be a safe integration site in many *Sphingomonas* strains, and also for other commonly-studied bacteria, and that other conserved sites for benign integration are more practical. We target such a benign integration site and reveal that it is conserved in *Sphingomonas* and tags there result in no detectable bacterial fitness penalties in growth media or on plants. Taking advantage of the conservation of this site, we use a tandem array of three guide RNAs to isolate unique, neutrally tagged strains directly out of a heterogenous population.

Results

Design of a minimalist mutation and tagging construct

We designed a DNA barcoding construct (Fig 1A, Supplementary Fig 1) for the smallest footprint and lowest metabolic cost. As an antibiotic resistance marker, necessary for efficient selection of transformed cells, we chose tetracycline resistance regulated by the native *tet*R repressor [15]. This system is broadly effective in *Sphingomonas* in our experience, and importantly is not actively transcribed in the absence of tetracycline, and so is therefore off in most experimental settings [16]. We flanked the tetracycline resistance gene with flippase (*frt*) sites that would allow it to be fully excised if desired, and did not include fluorescence proteins or any other constitutively expressed genes because these have limited uniqueness as well as potential metabolic costs. Downstream of the tetracycline resistance cassette we included two *rrnB* rho-independent transcriptional terminators to prevent transcription beyond the insertion. Following the terminator region is a unique 16 bp random sequence serving as a DNA barcode. The terminator and barcode are flanked by PCR priming sites, including not only unique priming sites but also the common 16S rRNA gene (rDNA) priming sites 515F and 799R/806R that amplify variable region 4 (V4) [17]. The full amplifiable region is ~300 bp, nearly identical in length to actual bacterial V4 16S rDNA amplicons and long enough to be purified efficiently with Solid Phase Reversible Immobilization (SPRI) beads. We modified the pSPIN pSL1142 vector from [8] to include this barcoding construct on the transposon (Methods), and further modified the vector backbone to include a dominant negative synthetic *rpsL* gene that could confer streptomycin sensitivity, and thus be used for negative selection in *Sphingomonas*

which are broadly naturally resistant to streptomycin [18]. The final vector library, loaded with random barcodes, we named pSPIN-LL.

Site-specific integration in diverse *Sphingomonas* strains

We chose 8 *Sphingomonas* from our sequenced culture collection that represented diverse lineages (Fig 1C-E) to gauge broad applicability in the genus, and prepared to transform each in both a neutral location and within a coding sequence (Fig 1B). For the neutral location, we reasoned that the widely-conserved and well-studied attTn7 site would be an ideal target because negative fitness effects are rarely reported. The Tn7 transposon recognizes a conserved attTn7 site in the final 36 bp of the glucosamine-6-phosphate synthetase (*glmS*) gene, and typically integrates about 25 bp downstream of the stop codon. Since the mini-transposon of the CRISPR RNA-guided integrase inserts around 50 bp downstream of the guide sequence, we only considered guide sequences beginning within the final 75 bp of *glmS* that would lead to insertions downstream of *glmS.* We looked for a site with minimal variation between all strains in our collection so that each guide would have utility in multiple strains, and identified six sites in *glmS* that had 5'-CN-3' PAM motifs in at least 97% of our strains. For each candidate site, we asked how many guide sequences would be required to target most of our culture collection. We discovered that a family of 10 guides covered 40% of the collection assuming perfect matches, and over 90% allowing for 2 mismatches (Fig 1F). Since the extent of conservation for each PAM site was similar, we arbitrarily chose one of the guide families with which to proceed, and checked that each guide in that family was unique within its target genome (Fig 1G). We note that due to a serendipitous error discovered later, the guide we designed for SphUPP:SBL2-17 differed from its actual genomic target by 5 bp (Fig 1G, Supplementary Fig 2).

For the coding sequence, we targeted the phytoene desaturase (*crtI*) gene (Fig 1B), which is a critical upstream step in the production of the carotenoids that give *Sphingomonas* their vivid yellow and orange colors [19]. Mutations in *crtI* provide a clear white colony phenotype that allows successful mutagenesis to be assessed without molecular methods [18]. We targeted conserved PAM sites in the first 100 bp of the coding sequence and found five sites conserved in 97% of strains. As with *glmS*, the family of guides at each potential guide was similarly effective in-silico (Fig 1F), so we chose a site and confirmed its uniqueness within each target genome (Fig 1G). For *crtI* also, the guide for SphUPP:SBL2-17 differed from its actual genomic target, in this case by 4 bp (Fig 1G, Supplementary Fig 2).

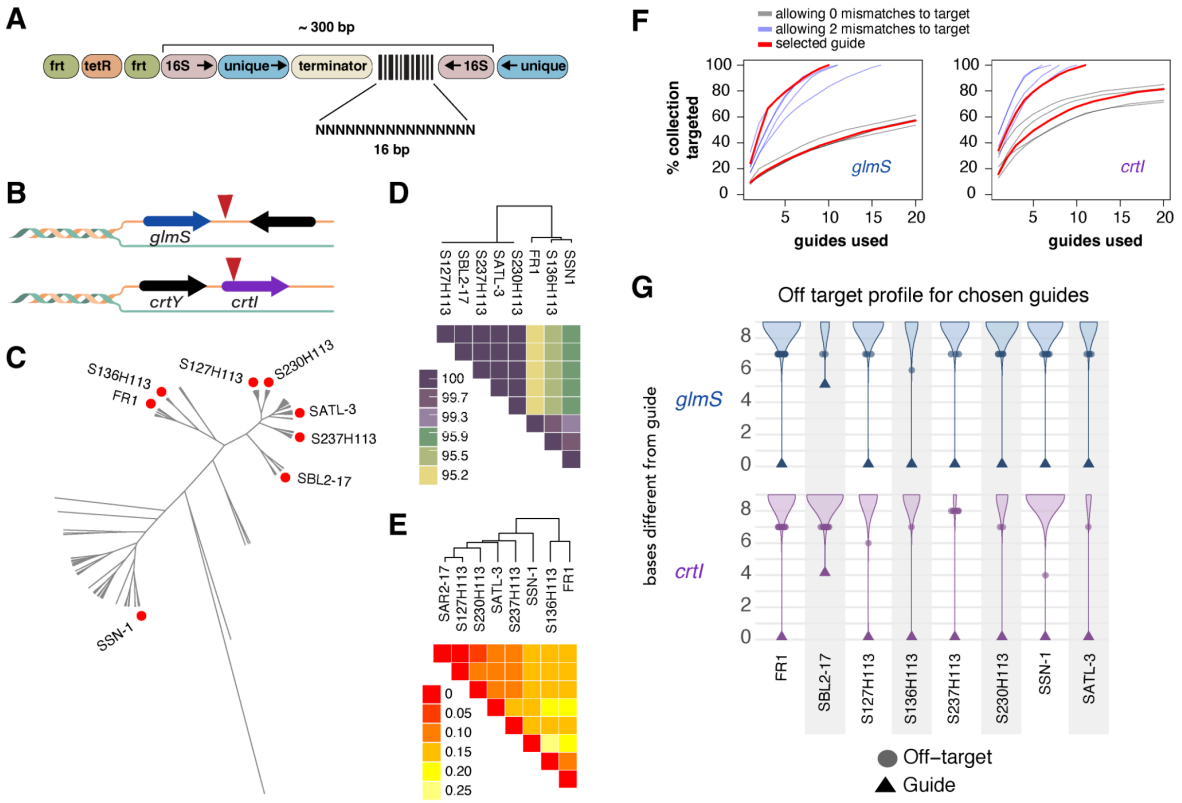


Figure 1. Design and planning of mutation constructs

A) Design of transposon payload for pSPIN-LL. frt, flippase sites; tetR, tetracycline resistance; unique, unique priming sites; 16S, 16S rDNA priming sites; N, random nucleotides. Arrows represent the direction of primer extension during PCR. **B)** Diagram indicating targeted region for neutral genome tagging (top, *glmS*) and for inducing functional mutations (bottom, *crtI*). **C)** Unrooted phylogenetic tree of 144 nonredundant (MASH distance > 0.01) *Sphingomonas* genomes, with positions of strains for which transformation was attempted indicated by a red dot. **D**) Comparison of percent nucleotide identity in the V4 16S rDNA sequences of the attempted strains. **E)** Comparison of MASH distance between whole genome sequences of the attempted strains. **F)** Number of strains from our culture collection (y-axis) that can in principle be cumulatively targeted by a family of up to 20 guide sequences (x-axis). Each grey or blue line represents a family of 20 sequence-related guides derived from a single conserved PAM site, with grey lines representing the case of perfect alignment to the culture collection, and blue lines allowing for alignments with up to 2 mismatches. The thicker red line represents the guide family chosen for subsequent experiments. **G)** Sequence similarity of off-target regions in each genome to the guide sequence designed for that genome as determined via BOWTIE2 alignment. Circles are plotted to represent discrete off-target events only for the most similar off-targets.

We synthesized all chosen guides (Supplementary Table 1 and Supplementary Table 2), ligated each into pSPIN-LL, and then conjugated pSPIN-LL into the appropriate *Sphingomonas* (Methods). We then plated on tetracycline plus streptomycin to select for transformants that had lost the pSPIN-LL plasmid. For both *glmS* and *crtI* mutants, we performed two colony PCRs: one to verify that the barcoded tagging construct was present, and a second to verify

that the pSPIN-LL plasmid could not be detected. We recovered at least one transformant in all 8 strains, demonstrating that integration is broadly possible in the genus. For subsequent analysis, we decided to focus our analysis on those five strains for which we recovered insertions resulting from both *crtI* and *glmS* guides. We attempted to map the insertion site(s) using transposon site insertion sequencing, but this was initially unsuccessful due to factors we subsequently optimized, so for expediency we selected a single candidate colony from each transformation and determined the transposon's location and barcode sequence using pooled whole genome sequencing with long reads (Supplementary Table 3, Fig 2A-B).

In all ten genome sequences from the five strains, the transposon was present as a single insertion in the same orientation. For the five *crtI* mutants, the insertion was in the coding sequence of *crtI*, and all transformants had lost their pigment (Fig 2C). For three of the *glmS*-targeting strains, the insertion occurred less than 200 bp downstream of the guide, while in the other two the insertion occurred thousands of bases away (Fig 2A). Remarkably, both the *crtI* and *glmS*-targeting insertions in SphUPP: SBL2-17 were within 100 bp of their target despite the mismatches in the guide, demonstrating that 85% nucleotide identity is sufficient if the target is unique!

Fitness of tagged strains relative to untagged counterparts

The fitness consequences of the *crtI* mutations are beyond the scope of this study. To determine if our *glmS*-targeting insertions were phenotypically neutral, we focused on *S. melonis* FR1, SphUPP:SBL2-17, and SphATUE:S127H113 because these had insertions close to the attTn7 site. While the distant off-target tags in the other two strains might also turn out to be benign, we would prefer to create a collection of tagged strains in a consistent genomic location where fitness effects would be, if not avoidable, more predictable. We performed competition assays by mixing each tagged strain with its untagged wildtype version in a 1:1 ratio and growing the mixture on LB liquid, LB agar, and on 2 week old *A. thaliana* rosettes. We collected LB liquid and agar samples after 1 day of rapid growth at 28° C, and collected Arabidopsis rosettes sampled after 1 week of growth (Methods). We then used qPCR to determine if the ratio of the strains changed as a result of the competition, which would indicate a fitness difference, and detected none (Supplementary Fig 3). However, we note that the signal to noise in our assay was not sufficient to detect slight fitness penalties, and additional samples and assays will be needed to rule out minor effects (Supplementary Discussion).

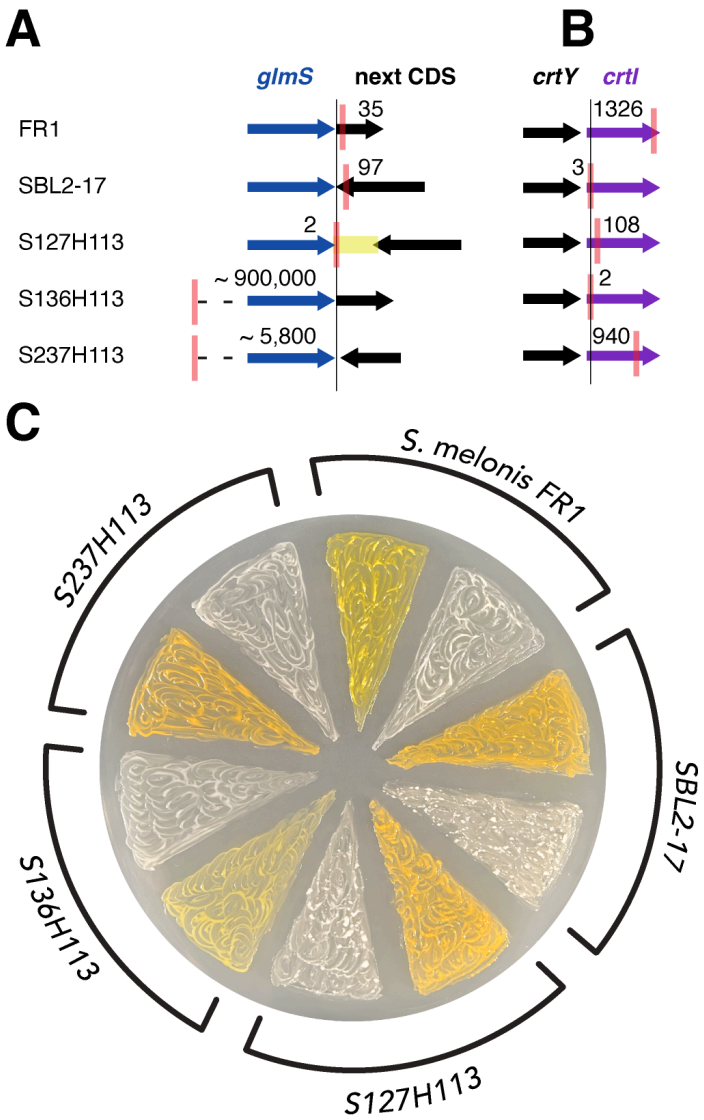


Figure 2. Confirmation of correct construct integrations

A) Diagram of mapped transposon insertions in the five strains, for the region downstream of *glmS*. The thin vertical line intersecting all rows represents the expected insertion location 50 bp downstream of the guide used, while the red vertical line within each row represents the actual observed location. The calculated distance between the observed and expected location in basepairs is indicated by an integer on the appropriate side of the expectation line. The yellow box for SphATUE:S127H113 represents deleted bases. CDS, coding sequence. **B)** Same as (A), for insertions in *crtI*. **C)** The *glmS*-targeting and *crtI*-targeting (pale color) transformants from each strain.

While mapping the insertions, we increasingly questioned whether the Tn7 site was truly a safe choice in *Sphingomonas*; ideally, the transposon should insert between two convergently terminating genes to avoid disrupting operons or promoters, but only three of our five strains had *glmS* convergently terminating with the next gene (Fig 2A), and one of those three (SphUPP:SBL2-17) had only 12 bp before the next predicted open reading frame (ORF) offering

an extremely small target. For the other two strains, the gene following *glmS* was co-directional and was 60 bp or less from the *glmS* stop codon, such that any insertion in the 3′ UTR would likely interfere with its transcription. While SphATUE:S127H113 represented an ideal case with the next convergently terminating ORF 697 bp downstream, unexpectedly our transposition created a deletion of about 700 bp in this strain such that the final bases of this ORF were eliminated (Fig 2A). Thus, all of our *glmS*-targeting strains had some sort of defect, but nonetheless appeared to grow normally.

Suitability of tags for amplicon-based quantification

Regardless of potential fitness consequences on living cells, all insertions should be similarly quantifiable by amplicon sequencing. To test this, we used both unique transposon-specific primers as well as V4 16S rDNA primers that not only amplify our construct (Fig 1A) but also each strains' native 16S rDNA gene. We first mixed DNA from the 10 strains in a balanced, equimolar ratio, and then created a stepwise linear mix and a stepwise exponential mix by manipulating the volume of each DNA stock relative to what was added to the balanced mix (Fig 3A). The starting balanced mix had slightly uneven abundances of the 10 barcodes, but both primer sets captured highly correlated barcode relative abundances for all mixes (Fig 3A, Fig. 3B). Amplicons also accurately captured the linear and exponential adjustments; we multiplied the observed balanced abundances by the linear or exponential factors to calculate expected relative abundances for each altered mix, and saw a high correlation with the observed abundances (R^2=0.96 and R^2=0.99 respectively, Fig 3C, Fig 3D).

The main value of using 16S rDNA primers in this context is to also detect uncharacterized bacteria in a complex sample. However, to be able to compare barcode abundances to native 16S rDNA abundances, it is important to know how much of each amplicon a single tagged strain contributes. We therefore used 16S rDNA primers to sequence amplicons made individually from each isolated strain (Fig 3E). Importantly, the sites to which the primers anneal are identical in all five of our strains and in our tagging construct. After correcting for 16S rDNA copy number in our strains by dividing by the known number of 16S rDNA copies derived from the closed genomes, it was clear that some variability existed between barcodes, including for different barcodes within the same strain (Fig 3F). Notably, barcodes for *S.melonis* FR1::*crtI* and Sph-ATUE:S237H113::*glmS* showed the lowest amplification efficiency relative to 16S rDNA in single strains, suggesting their underrepresentation in the balanced mix could be due in part to these same amplification inefficiencies (Fig 3A). For each barcode, we calculated a correction factor by dividing the copy number-normalized 16S rDNA abundance by the mean barcode abundance (Fig 3F).

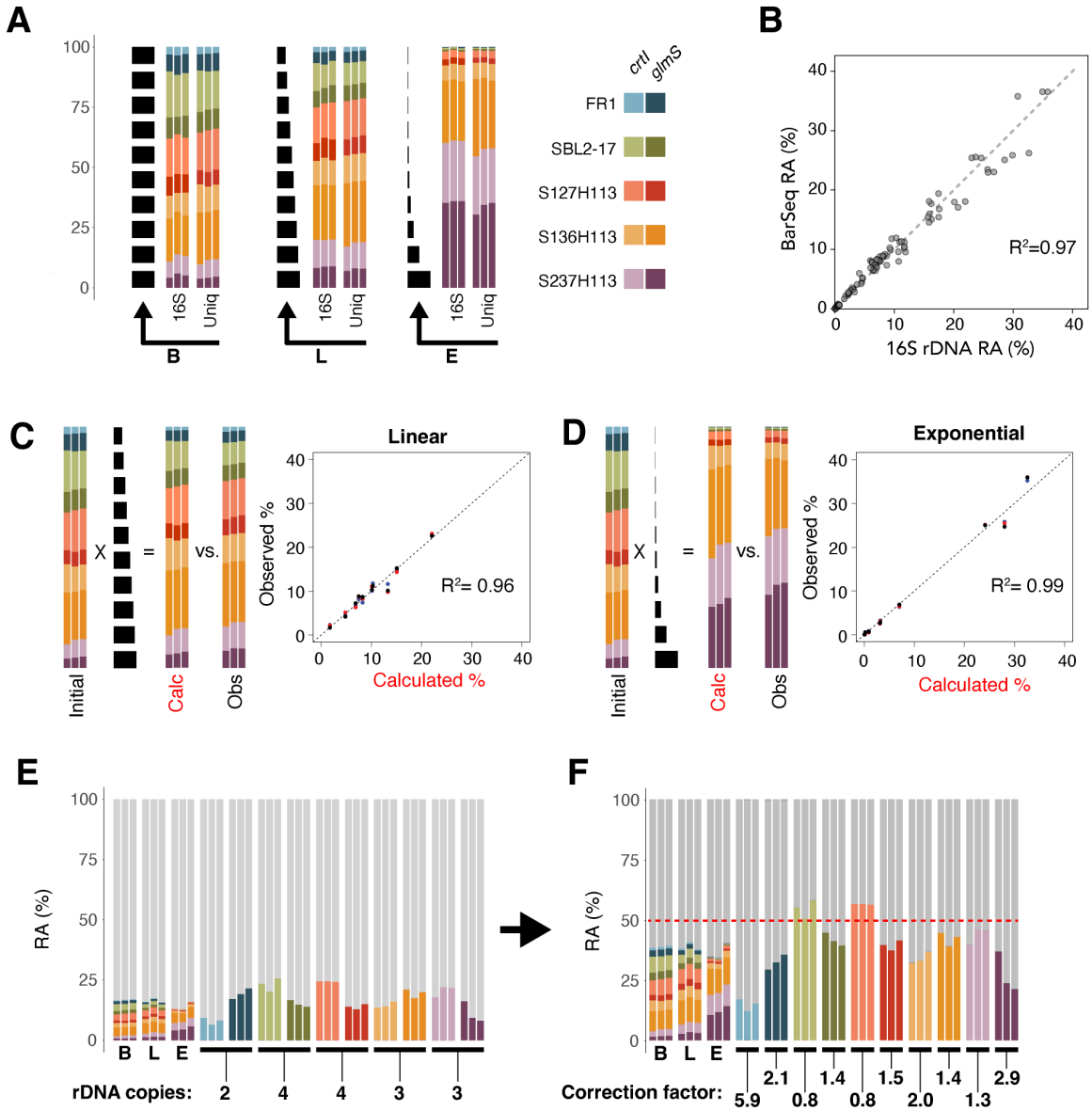


Figure 3. Quantitative validation using DNA mixtures

A) Relative abundances of barcode sequences in balanced (B), linearly-varied (L), and exponentially-varied (E) mixtures of the 10 tagged strains following the color legend at right, using either 16S rDNA (16S) or construct-unique primer sets (Uniq). Each stack of 10 black bars represents the relative contribution of each of the 10 strains to the pool. **B)** Pearson correlation between barcode abundances as detected by 16S rDNA primers vs. as detected by construct-unique primers, considering all DNA mixes. **C)** Comparison of relative abundances observed in the linearly-varied mix with those expected after applying the linear multiplication factors to data from the balanced mix. Cartoon diagram at left, Pearson correlation at right. **D)** Same as (C), but for the exponentially-varied mixture. **E)** Left: Relative abundances of amplicons of both total 16S rDNA (grey) and barcodes obtained with 16S rDNA primers for the three mixes as well as for individually-amplified strains. The 16S rDNA copy number as obtained from the closed genome sequence is shown below. **F)** Same as (E) but after dividing the 16S rDNA abundances by the 16S rDNA copy number in each strain. The red dotted line at 50% represents the expected

abundance of the barcode construct if it amplified with equal efficiency as 16S rDNA. The correction factor represents a constant by which barcode abundance can be multiplied to approximate normalized 16S rDNA abundance.

Quantification of tagged strains in complex environments

We inoculated our *glmS*-targeting tagged strains as a synthetic community on non-sterile *Arabidopsis thaliana* plants. For reasons of technical proof of concept, we decided to include the *glmS*-targeting SphATUE:S136H113::miniTn and SphATUE:S237H113::miniTn strains that had considerably off-target insertions, and note that the effect of those off-target insertions on those strains' biology is outside the scope of this work. In addition to inoculating the five strains (Mix5), we also created an inoculum of Mix5 mixed with a complex uncharacterized community cultured from river water that was enriched for *Sphingomonas* (WaterCom, Fig 4A). Interestingly, SphATUE:S136H113 was dominant in the Mix5 inoculum, despite the strains having been carefully mixed at equal OD_{600}, suggesting this strain absorbs less light at 600 nm (perhaps due to smaller cell size). Although three of the five strains in the inoculum share the same 16S rDNA sequence (Fig 1D), *S. melonis* FR1 and SphATUE:S136H113 were unique, allowing us to again calculate the efficiency of tagging construct amplification compared to 16S rDNA amplification for those strains, as we did previously in Fig 3F. The ratio detected this time was considerably different (Supplementary Fig 4), possibly a result of replication-associated gene dosage [20]. Although the 16S rDNA abundance in a genome is not a steady benchmark against which to evaluate copy number for other genes, we concluded that the barcode amplification efficiencies as calculated in Fig 3F were more accurate because those DNA preps were made from older, stationary phase cultures that were less likely to have partially-replicated genomes. We therefore used these values to correct barcode abundances; however, more rigorous copy number standards will be required to precisely account for barcode-to-barcode amplification variation.

The high density with which we inoculated plants led to an initial bacterial load that was several fold higher than the ultimate sustainable load that could be detected after 7 days (Fig 4B). The load of the barcoded strains alone also stabilized to a similar level even in the presence of the WaterCom (Fig 4C). While some bacterial families changed in abundance from day 0 to day 7, particularly an ASV classified as Enterobacteriaceae that was highly enriched at day 7 in WaterCom-inoculated samples (Fig 4A), our barcoded *Sphingomonas* maintained a similar fraction of the community. Moreover, the relative abundance of the barcoded strains among each other was essentially unchanged (Fig 4D), suggesting each had similar survival and growth rates in the phyllosphere. Importantly, the 16S rDNA sequences contributed by our tagged strains were common both on uninoculated plants and on plants inoculated only with the WaterCom (Fig 4A), and therefore would have prevented observing the true abundance of our strains if they were not tagged.

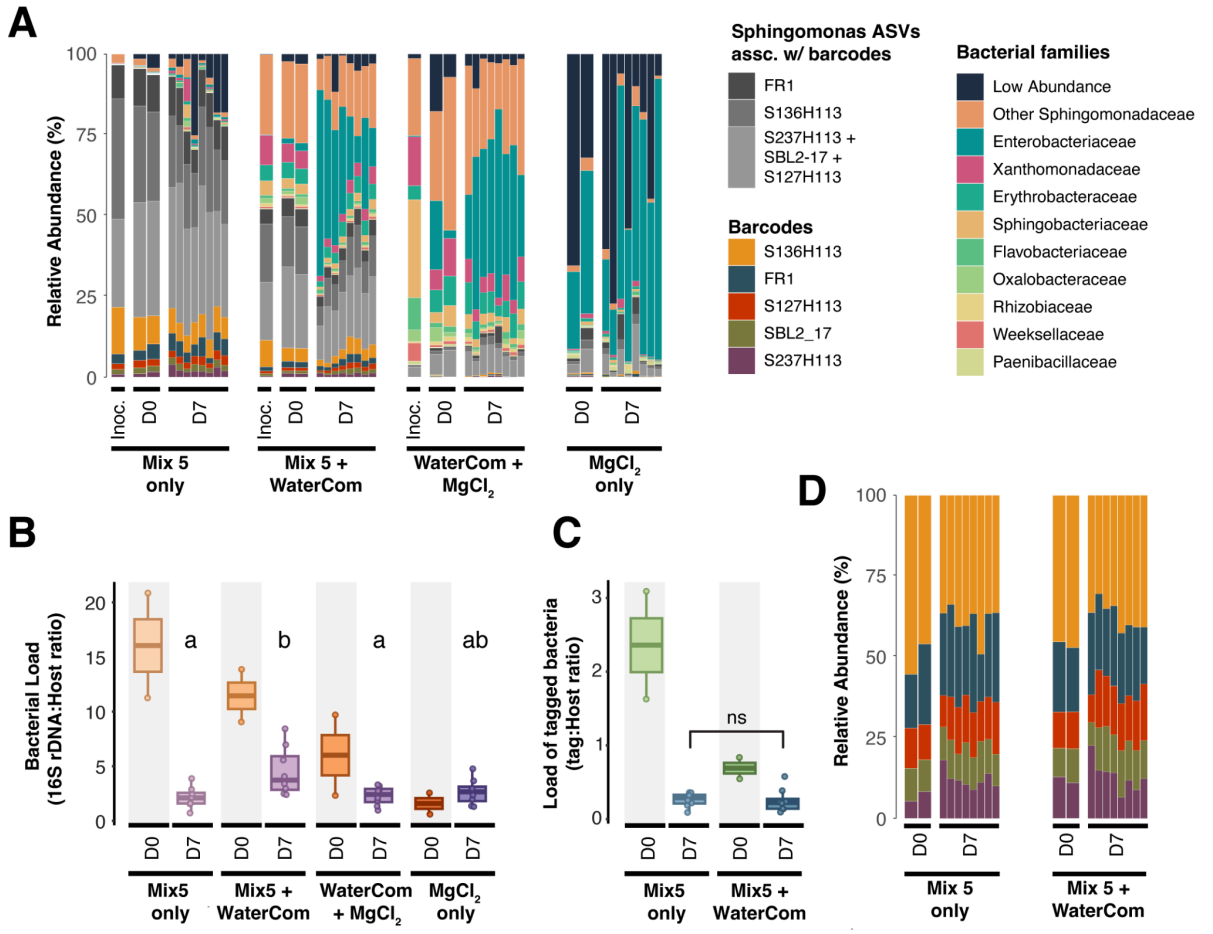


Figure 4. Tagged *Sphingomonas* colonization of plants

A) Relative abundances of 16S rDNA amplicons, including both barcodes of the tagging construct and 16S rDNA sequences, for different inoculum conditions. These included an balanced mix of only the barcoded strains (Mix5), a 1:1 mix of Mix5 and an uncharacterized bacterial community isolated from river water (WaterCom), a 1:1 mix of WaterCom with 10 mM $MgCl_2$ vehicle, and $MgCl_2$ only. The 16S rDNA sequences associated with the barcoded stains are shown in shades of grey (color legend), even if those sequences may have also been already present on the uninoculated plant or in WaterCom. Other bacterial ASVs are classified to the family level (color legend). Inoc, inoculum; D0, day 0 plant sample; D7, day 7 plant sample. **B)** Total bacterial load on plants, determined as the ratio of 16S rDNA (excluding tagging constructs) to the *Gigantea* gene using hamPCR. Statistical differences between groups at D7 were assessed using a Kruskal-Wallis test followed by Dunn's post hoc test with Benjamini-Hochberg correction for multiple comparisons. Different letters above the boxes indicate significant differences ($p < 0.05$). **C)** Same as (B), but for the load of tagging constructs (excluding all 16S rDNA). Statistical differences between groups at D7 were assessed using a Kruskal-Wallis test. No significant differences were detected ($p = 0.59$; ns). **D)** Relative abundances of the barcodes only for those conditions that included Mix5, following the color legend in (A). In all panels, barcode abundances shown were corrected for technical differences in amplification efficiency (Methods).

An improved universal tagging system in *Sphingomonas*

We were unsatisfied with the *glmS* 3′ UTR as a clean integration site in our five strains, due either to small landing pads intolerant of slight targeting errors, or to co-directional downstream genes for which transcription may be interrupted. To see if these features were more widespread in *Sphingomonas*, we investigated genomic structure around *glmS* in all non-redundant genomes in our culture database, ignoring highly-similar genomes differing by less than 1% MASH distance [21]. We recorded the orientation of the next coding sequence and its distance from the stop codon, and whether the 3′ UTR included any other non-coding feature (eg. tRNA). Indeed, in a majority (70%) of cases, the next gene or feature was either < 34 bp from the *glmS* stop coding and thus would be interrupted by transposons directed by our chosen guide family, or was co-directional (Fig 5A). We created an additional sequence resource from which to screen variation by culturing 250 diverse *Sphingomonas* from 25 geographically-separated Arabidopsis rosettes in Sweden (10 isolates per rosette) and sequencing the unindexed pool with long reads. We also saved pool aliquots as a glycerol stock for later bulk culturing (Methods). In this sequence resource as well, we annotated *glmS* and downstream features. As with genoms, in a majority of cases, insertion downstream of *glmS* appeared problematic (Fig 5A).

To search for a more robust insertion location, we adopted the targetFinder algorithm [22] to search 48 closed *Sphingomonas* genomes for convergently terminating ORFs, requiring a minimum gap size of 100 bp. We found that *rpoZ*, encoding DNA-directed RNA polymerase subunit omega, emerged as a gene flanking such a gap in a majority of strains. A full analysis of the gap revealed that in more than 90% of cases, both in non-redundant *Sphingomonas* genomes and the pooled mix of 250 diverse *Sphingomonas*, the *rpoZ* 3′ UTR provided an apparently safe landing pad (Fig 5B). We extended this analysis to the gamma-Proteobacteria *Pseudomonas* (Supplementary Table 4), another common member of the phyllosphere and one that is more extensively studied with Tn7 [9,23–25]. Surprisingly, the region downstream of *glmS* was also crowded in many genomes (Fig 5C). Unfortunately, *rpoZ* does not pose a solution in *Pseudomonas*, with the next gene co-directional and quite close in all observed cases (Fig 5D). It is likely that a safe genomic location superior to *glmS* could be found through a similar genus-wide search.

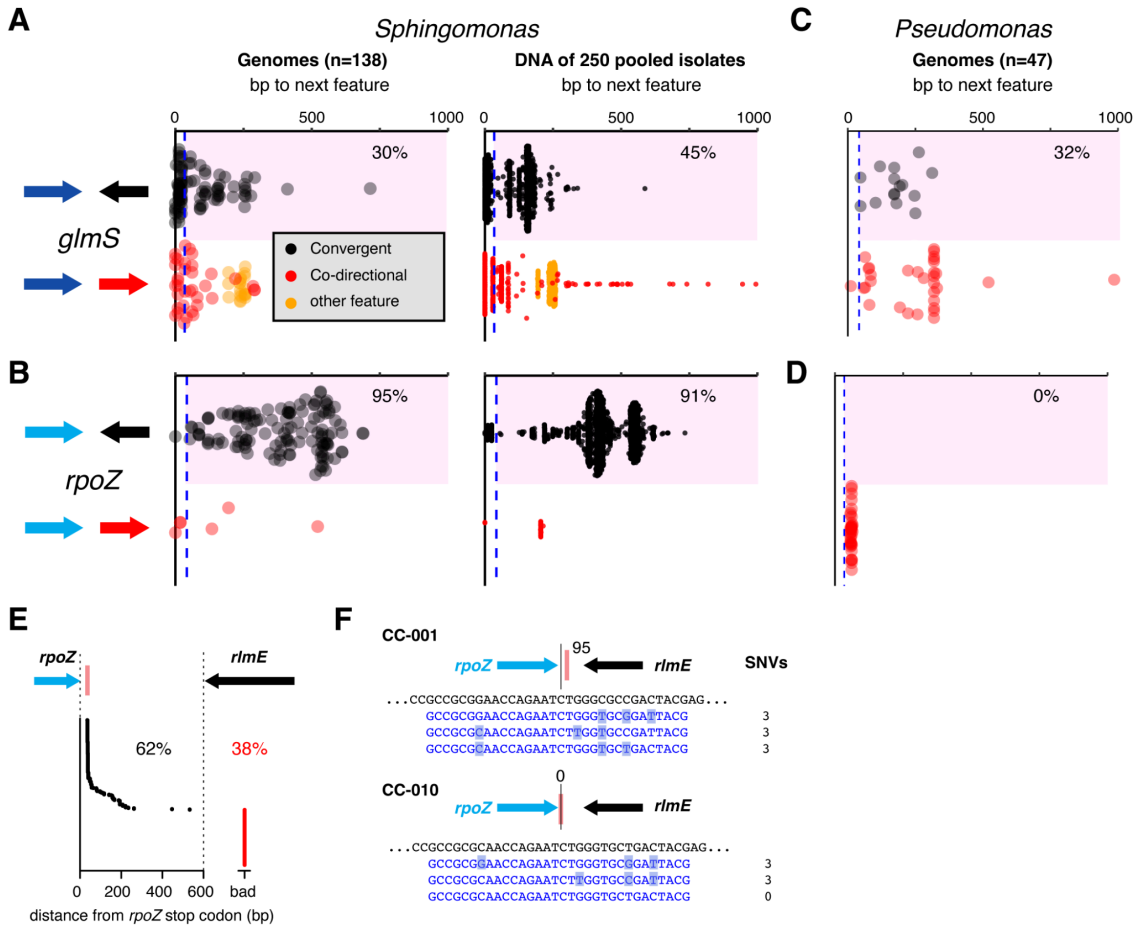


Figure 5. Integration loci for high-throughput tagging

A) Elements downstream of *glmS* in 138 non-redundant *Sphingomonas* genomes (left) or in a sequenced pool of 250 *Sphingomonas* isolates. The y-axis represents the *glmS* stop codon, and distance along the x-axis represents base pairs until a coding sequence that was convergent (black), co-directional (red), or until another type of annotated feature (orange). The dotted vertical line represents the expected transposon insertion location. The pink shaded box surrounds acceptable situations for which the convergently terminating ORF is downstream of the expected insertion. The percent of acceptable insertions is indicated. **B)** As in (A), but for the region downstream of the *rpoZ* gene. **C)** As in (A), but in a collection of *Pseudomonas* genomes. **D)** As in (B), but in *Pseudomonas*. **E)** tagIMseq mapping in a pool of 200 primary transformants with a guide targeting *rpoZ* in *S. melonis* FR1. The y-axis represents the *rpoZ* stop codon, and the distance to each point along the x-axis represents the distance to an observed transposon insertion. Black dots, and the black percentage, represent acceptable cases in which the insertion precedes the convergently terminating *rlmE* gene, while red dots and the red percentage represent insertions outside of this acceptable range. **F)** Single nucleotide variants between the genomes of isolates CC-001 and CC-010 (black) and the tandem triple *rpoZ* guides (blue) used to transform the pool of 250 *Sphingomonas*.

Encouraged by the conserved landing pad in the *rpoZ* 3′ UTR, we searched for a guide family in the final 75 bp of *rpoZ*, following the logic described earlier for *glmS* and *crtI* (Supplementary Fig 5). We then synthesized am *rpoZ*-targeting guide for *S. melonis* FR1, and

generated transformants using pSPIN-LL. To economically map the insertion locations and their associated barcode, we developed tagmentation transposon insertion site mapping by sequencing (tagIMseq), building on a method by [26], and validated tagIMseq on a template of pooled DNA from our 10 strains with insertions verified by genome sequencing, as well as on a template that was a suspension of whole bacterial cells from a single colony of one of those strains! (Methods, Supplementary Fig 6, Supplementary Fig 7). Equipped with tagIMseq, we pooled 200 *rpoZ*-targeting transformant colonies and mapped all insertions in the pool, which revealed that just over half of transformants were sufficiently on-target to fall in the acceptable landing pad (Fig 5E). To isolate correctly tagged colonies, we further improved our screening protocol, using initially use a colony PCR with an *rpoZ* specific primer and a transposon-specific primer to confirm the presence of a correct genomic insertion via band visualization in gel electrophoresis, and following that with tagIMseq on individual PCR-positive colonies; in all cases, PCR-positive colonies had a single on-target insertion in the *rpoZ* 3' UTR (Supplementary Fig 5).

We selected an on-target insertion and conducted a similar fitness test as we had used for *glmS*-targeting mutants, and detected no defects (Supplementary Fig 3). We therefore decided to establish feasibility for high-throughput direct tagging in this location by directly transforming our pool of 250 diverse *Sphingomonas* and directly isolating the tagged strains. To prepare the most broadly-useful guide sequences, we extracted *rpoZ* sequences from our pool of 250 strains, extracted all sequences in the guide family just validated in *S. melonis* FR1, and clustered them at 93% identity (allowing for 2 mismatches in the 32 bp guide) using USEARCH [27]. We then took the representative sequences from the top three most abundant bins and created a triple tandem guide sequence for ligation into pSPIN-LL (Supplementary Fig 8). Next, we transformed the pool of 250 strains directly from the recovered glycerol stock, and plated on selective media. From our transformants, we analyzed ten colonies by colony PCR to confirm an insertion downstream of *rpoZ*, found two that had a positive signal, and performed tagIMseq on those colonies to identify bases associated with all integration locations. In each case, this revealed a single insertion. We sequenced the genomes of these strains, which further revealed that in both cases the next coding sequence was a convergently terminating *rlmE* gene (Supplementary Fig 9). Through comparison of the genome sequences flanking the transposon with high quality raw long reads from the 250-member pool, we identified raw reads that were identical to the genomes except for the transposon insertion. These most likely represent the wild type strains that were transformed, and confirm insertion into a featureless gap with no bases deleted. The two recovered genomes differ from each other with a MASH distance of 0.22 corresponding to variation in approximately 1 in 5 bases [21], and while one of the genomes had an *rpoZ* sequence exactly matching one of the three guides, the other differed from all guides by 3 bases (Fig 5F).

Discussion

Here we have designed a minimalist tagging construct and targeted it to multiple genomic locations in diverse isolates of the genus *Sphingomonas.* Prior to the development of CRISPR RNA-guided intragrases, this would have required homologous recombination, a process which requires inclusion of bacterial genomic sequences on both sides of the insertion construct [18]. While tools using linear DNA and short (35-80 bp) regions of homology exist for efficient homologous recombination in *E.coli*, [28], homologous regions of 500 bp or more are typical for non-model organisms [18]. The longer the homologous sequence required, the less likely it is that the same construct will work on a genetically-distinct organism. This practical difficulty has driven the interest in Tn7, which uses a protein complex that recognizes sequence and structural features to target the conserved and widespread attTn7 site [29]. The ease of inserting DNA with Tn7 is such that, when faced with a situation for which an organism is missing an attTn7 site or if the native attTn7 site is not tolerated, researchers have sometimes taken the circuitous approach of creating a "landing pad" attTn7 site somewhere else via homologous recombination such that the innate Tn7 proteins can recognize it [30,31].

Like Tn7-targeting MoBacTags [6], our constructs include conserved 16S rDNA priming sites that allow direct comparison of barcode abundances to background 16S rDNA abundances in a complex sample. However, our 16S rDNA priming sites correspond to the V4 region instead of the V5-V7 region, and our amplicon is precisely size-matched to actual V4 16S rDNA amplicons. As with MoBacTags, we observed barcode-specific effects on the tag-to-16S read count ratios for different tags, despite each having the same priming sites and being in the same bacterial strain (Fig 3), supporting the conclusion that individual barcode sequences influence amplification efficiency [6,32]. Here we have followed the approach of Ordon et al. [6] to use the tag-to-16S ratio in DNA preps from individual strains to calculate a "correction factor" for the barcode abundance in each strain, but we suggest that this correction factor, while arguably better than nothing, is flawed. In our data, the tag-to-16S ratio differed between biologically-independent samples (Fig 3F, Fig 4A, Supplementary Fig 4). Further, it is well-documented that genes near the origin of replication, often including 16S rDNA, may exist in heightened copy number in rapidly growing cells [20]. Therefore, the ratio of 16S rDNA is expected to change depending on the growth conditions, and likely in ways that cannot be predicted or experimentally determined in complex conditions. We therefore propose an improved correction factor will require comparison of barcode construct amplification with that of a reference amplicon in close genetic proximity, where a 1:1 template ratio is guaranteed.

One of our insights from genomic analysis was that the Tn7 site may be more problematic as a fitness-neutral site than is generally assumed. While there are scattered reports of fitness penalties due to Tn7 integration [9,31], in most cases the downstream gene is never identified, nor is the fitness on the Tn7-tagged strain compared to the wild type [6]. It is encouraging that the *Sphingomonas* Leaf 257 tagged by Tn7 in [7] apparently had no fitness effects in vitro or in planta, because this genome also encodes a putative bi-functional diguanylate

cyclase-phosphodiesterase on the same strand as *glmS* only 78 bp downstream. While such genes downstream of *glmS* may sometimes be dispensible, this may not apply to all environments. The gene downstream of *glmS* in *S. melonis* FR1, for example, is a ROK family fructokinase. Such enzymes are important for fructose metabolism and its disruption might not slow growth in rich media, but it could be problematic in oligotrophic leaf environments where fructose is a primary carbon source [33]. In any case, it is clearly worth avoiding disrupting open reading frames and operons as a general rule. We discovered the region downstream of *rpoZ* is superior to that downstream of *glmS* in *Sphingomonas* genomes in our collection, but is much worse in *Pseudomonas*. This underscores the need to consider different safe sites depending on the genetic background, which unfortunately is not a flexibility of Tn7.

Our unintentional successful use of a guide with 5 mismatches (Fig 1F, Fig 2) revealed that some sequence variation is tolerated. While this could make it difficult to direct insertions in some locations within a genome, it is an advantage when targeting a unique location across multiple diverse genomes, and thus facilitates the study of natural variation in bacteria through genome tagging many strains. We exploited this tolerance for mismatches to design three guides for *rpoZ* to broadly target plant-associated *Sphingomonas* and used the guides to recovered unique, correctly-tagged isolates, including one that had three nucleotide variants distinguishing it from the nearest guide sequence (Fig 5, Supplementary Fig 9). This successful integration, plus the essentially on target integration in SphUPP:SBL2-17 (Fig 2A) involving 5 mismatches, suggest that designing our tandem guide with only 2 bp differences between each guide was too conservative, and that a more broadly useful guide array can be designed.

Unlike in *E. coli* where off-target insertions are rare exceptions [8], in *Sphingomonas* we recovered many off-target insertions. The fact that several of our insertions in *crtI*, and many of the insertions we mapped downstream of *rpoZ*, were beyond the expected 50 bp but still relatively close, demonstrates that proximal off-target insertions are common. Fortunately, they are often still acceptable for many purposes. Recently, a genetic screen for factors influencing VchCAST transposition uncovered activator and inhibitor genes that are distributed differently among Gram negative bacteria, and contribute to poor efficiency in some species [34]. The authors modified the VchCAST delivery vector to express bacteriophage λ-Red genes (exo, beta, and gam) and were able to increase efficiency; such a system is likely to further improve efficiency in *Sphingomonas*.

Due to the relatively high off-target rates, we worked to develop an extremely efficient screening and transposon mapping pipeline that, remarkably, can be done without a DNA prep on primary transformed colonies during their lifespan. First, a colony PCR enables recognition of probable on-target mutations, and second, colony tagIMseq paired with a mail-order rapid sequencing service enables precise mapping of transposon site insertions. This speed means that it is not necessary to invest effort in subculturing colonies or making DNA until a single on-target insertion is confirmed, removing a major hurdle in high-throughput tagging.

Methods

Creation of pSPIN-LL barcode library

As a transposon payload, the pSPIN-LL plasmid includes 1) a tetracycline resistance gene flanked by flippase (frt) sites, followed by 2) a transcription termination region and a unique 16 bp barcode flanked by both V4 16S rRNA priming sites and unique priming sites. We also modified the pSPIN backbone outside of the payload by adding a synthetic rpsL gene from pVHD [15] that serves as a negative selection marker in most *Sphingomonas* [18]. The plasmid library was constructed as described below.

– Construction of pSPIN-LL payload

Our construct design was largely ordered directly as a synthetic construct (Integrated DNA Technologies, Coralville, IA, USA), but also involved some cloning intermediates. We describe only an overview of the intermediate processes, as they did not proceed directly to our final construct and ultimately are not important because the final sequence of the construct is published here and the plasmid itself is available upon request. Briefly, our original synthetic fragment included a flippase site, a gentamicin resistance gene, a green fluorescent protein (GFP), another flippase site, and finally a region flanked by useful PCR sites that included transcription terminators and a 16 bp unique barcode. The barcode diversity was created by amplifying the construct with a custom oligo including random bases. We decided to replace the gentamicin and GFP in the original cassette with tetracycline resistance, and we therefore amplified the tetR promoter and tetR gene from pVHD [15] using primers L0129 and L0130, and amplified other parts of the construct we wished to keep and spliced them together into cloning vector pJET1.2 (Thermo Fisher Scientific, Waltham, MA, USA) using the GeneArt Gibson assembly kit (Invitrogen, Carlsbad, CA, USA). We collected hundreds of colonies, each representing unique barcodes, and re-isolated plasmid from the pooled colonies.

– Adding construct and *rpsL* to pSPIN

The synthetic *rpsL* gene from pAK405 conferring streptomycin sensitivity [18] was amplified using primers L0066 + L0067, and moved by Gibson cloning into a pSPIN backbone that we had been previously cut with FastDigest KpnI (Thermo Fisher Scientific, Waltham, MA, USA). This construct was then used as a template to amplify the region from the 3' end of the payload site through the newly-added *rpsL* gene, using the primer pair L0077+L0078. In parallel, we amplified the full construct from pJET1.2 using L0075+L0074. Finally, we cut again the original, unmodified pSPIN with FastDigest XhoI and Fastdigest KpnI and gel purified the vector. We finally cloned the insertion construct (L0075+L0074) and the plasmid *rpsL* fragment (L0077+L0078) into pSPIN using the Gibson method to make pSPIN-LL (Supplementary Fig 1), and selected transformants on 10 µg/mL tetracycline plates.

– Generating additional barcode diversity in pSPIN-LL

The first instance of pSPIN-LL construct had around 400 unique barcodes. To generate a more diverse library, we used pSPIN-LL as a template and re-amplified the majority of the payload using primers L0075+L0022, and we amplified the barcode and *rpsL* gene with L0261+L0078. These fragments were cloned by Gibson into original unmodified pSPIN that had been cut with both XhoI and KpnI and gel purified, thus remaking pSPIN-LL such that each colony represents a unique barcode.

Checking guide RNAs for off-target risks

Annotation-independent off-target analysis was performed using Bowtie2 (v2.5.4) within the framework established by CAST-guide-RNA-tool (https://github.com/sternberglab/CAST-guide-RNA-tool), with slight modifications. While the same computational environment was used for the analysis, a customized Bowtie2 search strategy was used of end-to-end alignment to verify the full 32 bp gRNA sequence, a 5 bp seed length (-L=5), zero allowable mismatches in the seed region (-N=0), and increased search depth (-D=10). Potential off-targets were identified through direct quantification of mismatch counts in the resulting SAM files and visualization performed in RStudio using the ggplot2 [35] package.

Plasmid DNA Extraction

Plasmids were extracted from *E.coli* by alkaline lysis [36]. For cloning into pSPIN-LL an additional chloroform purification was performed by adding 400 µL chloroform to the resuspended plasmid DNA, vortexing, and mixing the clarified aqueous phase with 1/9 volume 3M Sodium Acetate (NaAc) and 8/9 volume of isopropanol to form a precipitate collected by centrifugation. This pellet was washed with 80% ethanol, and finally resuspended in 500 µL of elution buffer (EB; 10 mM Tris pH 8.0). Purified plasmid DNA was treated with 1µL / 100 µL of RNase A (10 mg/mL, Thermo Fisher Scientific, Waltham, MA, USA), and further cleaned with solid-phase reversible immobilization (SPRI) beads [37] using a 1:1 ratio of beads to DNA. DNA concentration was quantified using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions.

Inserting guides into pSPIN-LL

Cloning of guide sequences into pSPIN-LL plasmids was performed as described in [38]. The 32 bp target sequence for the CRISPR integrase plus the 2 bp Protospacer adjacent motif (PAM) site was ordered as two complementary oligos that were phosphorylated with T4 Polynucleotide Kinase (Thermo Fisher Scientific, Waltham, MA, USA) to enable ligation and annealed to make a duplex. Primer names used for oligo duplex are detailed in Supplementary Table 2. All primer sequences can be found in Supplementary Table 1.

Confirmation of correct guide insertions in *E. coli*

Colonies of transformed *E. coli* that grew on tetracycline selective media were screened by colony PCR using one primer specific to the guide sequence (the one starting with 5'TTCAC…) and another specific to the transposon construct (L0056). See Supplementary Table 1-2

Transformation of *Sphingomonas*

The *E. coli* diaminopimelic acid (DAP) auxotrophic donor strain WM3064 was made electrocompetent as in [39]. A ligated pSPIN-LL plasmid including a guide sequence and the randomly barcoded construct were introduced by pipetting 5 µL of the resuspended ligation reaction into 50 µL of chilled electrocompetent *E. coli* in 1 mm cuvettes. The mixture was pulsed with standard conditions for *E. coli (*25 µF, 200 ohm, 1800 V) using the Gene Pulser Xell (Bio-Rad Laboratories, Hercules, CA, USA). Following the recovery period, the *E. coli* cells were pelleted by centrifugation, resuspended in 60 µL of LB + 0.3 mM DAP (Thermo Fisher Scientific, Waltham, MA, USA) liquid medium, and plated on LB agar plates containing 10 µg/mL tetracycline. The plates were incubated overnight at 37°C, allowing only the bacteria carrying the modified plasmids to grow.

The plasmids were conjugated from the above transformed *E. coli* cells to *Sphingomonas* through a direct cell-to-cell contact mechanism during an overnight co-culturing step. *E. coli* cells carrying the modified pSPIN-LL vectors were grown overnight at 37°C in LB media with 0.3 mM DAP and 10 µg/mL tetracycline. When they reached an OD600 of 1.0, 3 mL of bacterial culture was pelleted, washed via resuspension in fresh LB liquid supplemented with DAP, and resuspended in 500 µL LB media. Similarly, the *Sphingomonas* strains were grown at 28 °C for 2 days in LB media with 100 µg/mL streptomycin, and 3 mL was washed and resuspended in 500 µL of LB. The *E. coli* and *Sphingomonas* were combined, and this 1 mL was again pelleted, resuspended in 60 µL LB liquid with DAP, pipetted as dot on LB agar with DAP but no antibiotic, and incubated at 28°C for 16 hours.

The conjugation mix was then scraped off with a sterile loop, plated onto LB Agar plates with 10 µg/mL tetracycline and 100 µg/mL streptomycin and without DAP to remove *E. coli* cells and select for transformed *Sphingomonas* without pSPIN-LL.

Confirmation of transformed *Sphingomonas*

By colony PCR

Initially, colony PCR was performed only to rule out the continued presence of the pSPIN-LL plasmid using primers L0066 + L0067 specific to the synthetic *rpsL* construct, and to verify the presence of the transposon insertion using unique primers L0116 and L0262. Later, only a primer binding to the guide location (eg. L0427) and a primer binding to the 5' end of the

insertion construct (L0120) was used to reveal colonies with insertions within close range of their insertion.

By whole genome sequencing

For each of our five focal *Sphingomonas* strains, DNA from one strain each for the *crtI* and *glmS*-targeting transformations was pooled, and bacterial genome sequencing on the pool of 2 was performed by Plasmidsaurus (Plasmidsaurus Inc., South San Francisco, CA, USA) using Oxford Nanopore Technology with custom analysis and annotation. Reads containing the barcoded transposon construct were identified in the raw fastq file using the following regular expression and its reverse complement. Genomic regions just upstream and downstream of the transposon border were used to confirm that each barcode was associated with a single insertion, and to map that location.

```
"GCCTTTTTGCGT[ACGT]{16}TTACCGCGGCTGCTG"
```

By Tagmentation transposon Insertion site Mapping sequencing (tagIMseq)

Briefly, our method uses Tn5-based tagmentation followed by PCR using an insertion-specific primer and an adapter-specific primer, and is adapted with some additional modifications from [26].

Preparation of assembled Tn5

Tn5 was procured from Protein Production Sweden. Duplexes were prepared by annealing the appropriate oligonucleotides (listed below) in annealing buffer containing 100mM Tris (pH 8.0) and 500 mM NaCl. Duplex A was generated by mixing 9 µL of 100µM Tn5ME_A (L0361) with 9µL of 100µM aminoME_rev (L0376) and 2 µL annealing buffer. Duplex B was prepared similarly using Tn5ME_longB (L0416) and aminoME_rev (L0376). Oligonucleotides were annealed by incubation at 98°C for 3 minutes, followed by cooling to 23°C at a rate of 0.1°C /sec. Duplex A and B were then mixed. Tn5 stock was diluted by mixing 1 µL of stock with 16 µL of dilution buffer containing 50mM Tris-HCl (pH 7.5), 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 0.1% Triton X-100, and 50% glycerol. To the diluted Tn5, 3 µL of the mixed duplexes were added, and the mixture was incubated at room temperature for 30 minutes.

Tagmentation and initial PCR amplification

Both the tagmentation and amplification reactions were performed in a Bio-Rad T100 thermal Cycler (Bio-Rad Laboratories, Hercules, CA, USA). For DNA templates, the tagmentation mixture included 1 µL of DNA (10 ng/µL), 0.2 µL of assembled Tn5, and 2 µL of 5x TAPS-DMF-MgCl [40][2], and 6.8 µL nuclease-free water for a final volume of 10 µL. For bacterial colony templates, the colony was resuspended in 14 µL water and tagmentation was performed with 7 µL of this resuspension as template, 1 µL assembled Tn5, and 2 µL 5x TAPS-DMF-$MgCl_2$. Tagmentation mixtures were incubated at 55°C for 7 minutes, followed by Tn5 inactivation with 2.5 µL of 0.2% SDS and incubation at 55°C for an additional 7 minutes. A

first round of PCR was performed using primers annealing to the insertion construct just upstream of the barcode (L0426) and to the transposon-added Nextera overhang (L0417). To denature DNA just prior to PCR amplification and prevent "gap filling" at the Tn5-added adapters, 1.5 µL tagmentation product was mixed with 8.5µL nuclease-free water and denatured at 98°C for 3 minutes. The PCR reagents were then added to a final volume of 25 µL, including 0.5 µM each primer, 5x HF buffer, 200 µM dNTPs, and 0.25 µL Phusion™ Hot Start II DNA Polymerases (Thermo Fisher Scientific, Waltham, MA, USA). The PCR program was set to 98°C for 30 seconds, 30 cycles of 98°C for 30 seconds, 68°C for 15 seconds, and 72°C for 50 seconds, and finally a single extension at 72°C for 5 minutes. PCR products were purified using 0.6x SPRI beads and resuspended in 15 µL nuclease-free water.

Preparation of library

A second round of PCR was performed using 1 µL of the purified first-round PCR product as template and was amplified with a Nextera P7 index primer (Supplementary Table 1) binding to the overhang from primer L0417, and a TruSeq index primer (Supplementary Table 1) containing the P5 Illumina adaptor, with each at a final concentration of 0.5 µM. The reaction also contained 5×HF buffer, 200 µM dNTPs, 1.25 µL Phusion DNA Polymerase (Protein Production Sweden, working stock: 60 µg/mL; final concentration: 3 µg/mL), and nuclease-free water to a final volume of 25 µL. The PCR program was set to 98°C for 3 minutes, 18 cycles of 98°C for 30 seconds, 68°C for 15 seconds, and 72°C for 50 seconds, with a final extension at 72°C for 5 minutes. The final PCR products were purified using a double-sided SPRI bead cleanup (0.4x-0.6x), and then DNA concentration was quantified using Qubit.

Read processing and filtering

R1 reads were concatenated with the reverse-complement of the corresponding R2 reads to generate a single artificial continuous sequence. The combined reads were first filtered to keep only those containing the insertion sequence (GCAGGACGCCCGCCATAAA). (grep -Eo "GCAGGACGCCCGCCATAAA.*"). The second filter came with filtering the remaining reads that carried a 17 bp sequence (GGATGGCCTTTTTGCGT) before the insertion barcode, a 19 bp sequence (TTACCGCGGCTGCTGGCAC) after the barcode, and a 20 bp sequence (ACTTTATGGTTGCATCAACA) at the end of the insertion using grep -Eo "GGATGGCCTTTTTGCGT.{16}TTACCGCGGCTGCTGGCAC.*ACTTTATGGTTGCATCAACA". In addition, filtered reads were required to contain at least 50 bp of sequence downstream of the insertion end. After filtering, all sequences upstream of and including GGATGGCCTTTTTGCGT, as well as downstream of and including TTACCGCGGCTGCTGGCAC, were excluded to extract the insertion barcode and quantify barcode counts. Reads corresponding to the correct barcodes were then used to extract the 50 bp sequence downstream of the insertion. These potential downstream sequences were arranged from high to low by percentage for each barcode, and the highest one was selected for mapping to the reference genome.

Template preparation

For re-analyzing the ten *crtI* and *glmS*-targeting mutants as a pool, individual DNAs from each strain were combined. For analyzing *rpoZ*-targeting transformants as pools, colonies from the initial transformation were subcultured once by plating in a grid on a fresh streptomycin and tetracycline plate. After 24-48 hours of growth, all bacteria from the grid were pooled and the pellet was used for DNA extraction as described below. For analyzing individual bacteria, high quality DNA was either prepared from individual isolates, or, for colony tagIMseq, a single colony was resuspended in 14 µL of water and 7 µL was used as template.

Bacterial and plant DNA extraction

For bacterial pellets, the final pellets were resuspended in at least 750 µL of DNA lysis buffer containing 10 mM Tris pH 8.0, 10 mM EDTA, 100 mM NaCl, and 1.5% SDS. Especially large pellets (greater than 5 mm from the bottom of the tube) were suspended in proportionally greater volumes of buffer to ensure an efficient lysis, and 750 µL was used for lysis. The suspension was pipetted to a screw cap tube containing a 0.5 mL of a 1:1 mixture of 0.1 mm and 0.5 mm zirconia/silica beads (Bio Spec Products Inc., Bartlesville, USA) and homogenized in a FastPrep 24 (MP Biomedicals, LLC, Solon, OH, USA)instrument at 6 m/s for 1 min, following a subsequent incubation at 55°C for 1 hour. The tubes were centrifuged at 10,000 x g for 5 min, and the supernatant (about 600 µL) was mixed with 200 µL sterile 5 M potassium acetate in a 1.5 mL tube to precipitate the SDS. The tubes were chilled on ice for 10 min, then spun at 14,000 x g for 8 min, and the supernatant was transferred to a new 1.5 mL tube. Then, 360 µL SPRI beads were added to 600 µL of the supernatant (0.6:1 ratio). After mixing and incubating on a magnetic rack, the beads were cleaned with 80% ethanol twice, and DNA was eluted in at least 100 µL EB.

For pure bacteria cultures used to generate closed genomes, an additional RNAse and chloroform cleanup step was added. To the eluted DNA, 1 µL of RNAse A (Thermo Fisher Scientific, Waltham, MA, USA) was added and the mixture was incubated at 37°C for 30 minutes. The DNA was diluted in water to 400 µL, and 400 µL of chloroform was added. The aqueous phase was collected and the DNA was precipitated by mixing it with an equal volume of a solution containing 1 part 3M NaAc and 8 parts isopropanol. The precipitated DNA was pelleted by centrifugation at 14,000 x g for 5 minutes. The pellet was washed twice in 70% ethanol and eluted again in at least 100 µL EB.

For plant material, leaves were added to screw-cap tubes containing 0.5 mL of a 1:1 mixture of 0.1 mm and 0.5 mm zirconia/silica beads (BioSpec Products, Inc., Bartlesville, OK, USA), flash frozen in liquid nitrogen, and homogenized in the FastPrep 24 at 4 m/s for 15 s. Then, 750 µL lysis buffer was added and the extraction proceeded as above.

Pool of 250 *Sphingomonas* cultivation and sequencing

The 250-member *Sphingomonas* pool was produced by taking 25 *A. thaliana* individuals from around Uppsala, Sweden. Rosettes were rinsed in sterile water to remove adhered soil, and crushed in sterile mortar and pestle. The plant homogenates were resuspended in 0.5X phosphate buffered saline (PBS), and mixed with glycerol (20% v/v) to make culturable freezer stocks. Bacteria were recovered from glycerol by plating on low salt LB media plates with streptomycin (100 µg/mL) and cycloheximide (100 µg/mL). After 72 hours of incubation at room temperature, 10 randomly-chosen colonies from each Arabidopsis individual were picked and passed to a fresh LB plate for even growing, resulting in 250 isolates growing in parallel. After 72 hours of incubation, the 250 isolates were scraped off and homogenized in 10 mM $MgCl_2$. The bacterial suspension was adjusted to OD_{600} = 1, mixed with glycerol (20% v/v) and stored at -70°C. An aliquot of the resuspended mix centrifuged and the pellet used for DNA extraction as described above for bacterial pellets. The 250 isolates community DNA was sequenced with long reads on half of a PacBio HiFi Revio flow cell (Pacific Biosciences, Menlo Park, CA, USA) to generate approximately 45 Gb (Novogene, Beijing, China).

Amplicon libraries

Amplification of the uniquely barcoded region was performed in two rounds of PCR. In the first round, construct-specific primers (L0116 and L0262) were used. PCR reactions were performed in a total volume of 25 µL containing 1× Phusion HF buffer, 0.5 µL of 10 mM dNTP mix, 1.25 µL of each 5 µM primer, 1.25 µL of Phusion DNA polymerase (Protein Production Sweden, working stock: 60 µg/mL; final concentration: 3 µg/mL), 1 µL of template DNA (20 - 60 ng), and nuclease-free water to volume. Between the first and second PCR rounds, amplicons were purified using SPRI beads at a 1:1.1 DNA to SPRI ratio and eluted in 12 µL of nuclease-free water. Subsequently, 10 µL of the purified product was used as a template for the second-round of PCR, which was performed using the same reaction composition with Nextera index primers (Supplementary Table 1). The first-round PCR was performed for 15 cycles, and the second-round PCR for 20 cycles, resulting in a total of 35 amplification cycles. Amplification using 16S rDNA primers was performed identically, except that primers L0009 and L0010 were used for the first round of PCR. For the inoculation experiment, hamPCR was performed similarly to [37]. Primers L0009 and L0010 were used to tag the 16S rDNA region, and primers L0011 and L0012 were used to tag the *Gigantea* gene. Each reaction contained 5 µL of DNA template (25-250 ng), and the protocol consisted of an initial tagging PCR of 10 cycles, followed by bead cleanup and a second exponential PCR of 25 cycles using Nextera index primers.

Following amplification, reactions for each amplicon type were run on a 1% agarose gel to verify a distinct band and for rough quantification of product. Samples were then pooled based on the relative brightness of each band on the gel, as estimated by eye. The final pool was cleaned with SPRI beads with a 1:0.9 DNA to SPRI ratio to remove primer dimers. In the case of hamPCR, the final pool was run on a 2% agarose gel to separate the *Gigantea* and 16S rDNA

bands. Both bands were cut out, gel purified with the GeneJET Gel Extraction Kit ((Thermo Fisher Scientific, Waltham, MA, USA), and the GI product was mixed with the 16S rDNA product in a 1:4 ratio to reduce the relative contribution of host amplicon to the pool.

Genomes and annotation

Bacterial genomes used in this study and their source are listed in Supplementary Table 4. All bacterial genomes were annotated with BAKTA version 1.11.4 [41] using the full database of AA & DNA sequences and HMM & covariance models hosted at https://zenodo.org/records/14916843.

Benign insertion site identification

The TargetFinder [22] algorithm was used to identify gaps between convergently transcribed genes in 50 closed *Sphingomonas* genomes from our collection, requiring a minimum gap size of 100 bp. Genes flanking convergent gaps in multiple genomes were ranked to identify the most highly conserved genes associated with gaps.

Short read sequencing and amplicon processing

Different sub-libraries were first diluted to 6 nM, pooled by volume, and then sequenced at 80 pmol with ~5% PhiX on an Illumina MiSeq i100 series (Illumina, San Diego, CA, USA) using a 25M reagent kit with 300 cycles. Sequencing reads were demultiplexed by the Illumina platform, and only forward reads were used for downstream analysis. Primer sequences were identified and trimmed from their respective datasets using USEARCH v11.0.667 [27], including tagging construct specific primers, primers for the bacterial 16S V4 region, and plant *Gigantea* primers [37]. Reads that did not match the expected primer sequences were discarded. The remaining reads were then quality-filtered with USEARCH to retain only high-quality sequences. All filtered reads were pooled and dereplicated at 100% sequence identity using VSEARCH v2.25.0 [42]. Amplicon sequence variants (ASVs) were inferred as zero-radius operational taxonomic units (ZOTUs) using the UNOISE3 algorithm, which performs both error correction and chimera removal [43]. Taxonomic classification of 16S rDNA ZOTUs was carried out using the SINTAX classifier against the RDP 16S training set v18 (21k sequences) [44,45]. Finally, read counts were mapped to the ZOTUs at a 99% sequence identity threshold using VSEARCH to generate the final abundance tables. For the pure-culture calibration dataset, ZOTU counts of endogenous 16S rDNA were normalized to their known genomic copy numbers (Fig 3E). Ratios between copy-number–corrected endogenous 16S rDNA ZOTUs and synthetic tagging construct ZOTUs were used to calculate strain-specific amplification correction factors (Fig 3F). Finally, these correction factors were applied to the hamPCR dataset (Fig 4A) to account for barcode amplification bias before final relative abundance estimation. Bacterial load was calculated as the ratio of 16S rDNA abundance to Gigantea amplicons as in [37]. All

downstream normalization, aggregation, and visualization steps were performed in R using the packages reshape [46] and ggplot2 [35].

Tracking of tagged strains in complex plant environments

To evaluate the tracking efficiency of tagged bacterial strains within complex microbial environments, plant experiments were performed using a synthetic community (Mix5) consisting of five *glmS*-targeting strains: *S. melonis* FR1, SphUPP:SBL2-17, SphATUE:S127H113, SphATUE:S136H113, and SphATUE:S237H113. Each bacterium was grown separately in 50 mL of low-salt LB liquid media in a 200 mL baffled flask until reaching a $OD_{600}$of 0.5 - 1.0, with tetracycline (10 µg/mL) and streptomycin (100 µg/mL). Each culture was then centrifuged at 5000 x g and washed twice in 10 mM $MgCl_2$ to remove antibiotics, and finally resuspended in 10 mM $MgCl_2$ and adjusted to OD_{600} = 1. An uncharacterized community of primarily *Sphingomonas* (WaterCom) was isolated from a water sample from the Fyris river (Uppsala, Sweden) on solid LB media with streptomycin (100 µg/mL) and cycloheximide (100 µg/mL) incubated for 7 days. Colonies were washed from the plate with 10 mM $MgCl_2$, washed twice via centrifugation and resuspension to remove antibiotics, and adjusted to OD_{600} = 1. Four inoculum conditions were prepared to assess strain persistence and competitive ability: Mix5 alone; a 1:1 ratio of Mix5 to WaterCom; a 1:1 ratio of WaterCom to 10 mM $MgCl_2$; and a 10 mM $MgCl_2$-only negative control . To establish a baseline for community composition, 1 mL of each inoculum mixture was collected before inoculation as "Time-0" samples and pelleted for DNA extraction. Ten individual 3-week old *Arabidopsis thaliana* Col-0 plants were inoculated per condition using a 9cc airbrush paintgun (Biltema, Partille, Sweden) distributing 1mL of inoculum mixture between the ten individuals until the leaf surfaces were uniformly sprayed in two applications allowing them to dry in between. All leaves from two of the inoculated plants were harvested immediately following inoculation as "Time-0 plant" samples. The remaining eight plants per condition were grown for 7 days prior to final harvest. To ensure similar biomass as the Time-0 plant samples, only two mid-sized leaves were collected from each plant at the final harvest. Microbial load on plants was determined using hamPCR [37].

qPCR to determine fitness penalties due to tagging

To quantify both tagged and untagged bacteria, primers annealing to conserved loci in the single copy rpoZ gene were developed (primers L0443 + L0444) , producing a 108 bp amplicon. To quantify only tagged bacteria, primers annealing to the tetracycline resistance gene in our tagging construct were developed (primers L0453 + L0544), producing a 143 bp amplicon. Briefly, each DNA sample containing a mixture of wildtype and tagged bacteria was then amplified in three technical replicates per primer pair, using Taq Univer SYBR Green Supermix (Bio-Rad Laboratories, Hercules, CA, USA) on a Bio-Rad CFX Opus 384 machine. On the same plate, a log10 dilution series of tagged *S. melonis* FR1 *crtI*::miniTn, for which the ratio of tagging construct to *rpoZ* is known to be 1:1, was also amplified. Linear regression was then

used on the log10 dilution series for each primer pair to create a standard curve linking Ct value to dilution value. The mean Ct value for the three technical replicates in each sample was interpreted with the correct regression to determine an absolute dilution quantity of each template in each sample. Finally, the absolute quantity of tagging construct was divided by the absolute quantity of *rpoZ*. Each 10 µL qPCR reaction contained 5 µL SYBR mix, 0.25 µL of a working stock (10 µM) of each primer, and 1.5 µL DNA. To reduce the presence of *rpoZ* primer homodimers during reaction setup, the working stock was incubated at 95°C for 2 minutes and cooled on ice to prevent correct annealing, and the reaction was set up on ice and placed into the qPCR machine after it had already reached 95°C.

Data availability

All sequencing data have been deposited in the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena). They can be accessed under project number PRJEB108202. Scripts associated with this work can be found at https://github.com/LundbergLand

Author contributions

EMB, HB, TY, MT, and DSL conceived the study. EMB, HB, MT, NN, SH, LCH, and DSL performed inoculation and fitness experiments. EMB, MT, NN, and DSL created mutants. TY developed tagIMseq. EMB, HB, TY, and DSL analyzed data. EMB and DSL wrote the paper with input from all authors.

Acknowledgements

We thank Kyle Gervers and Fransiska Liesecke (SLU) for discussions. We thank Sophia Swartz (UC Berkeley) for discussions about Cas Tn machinery. *E. coli* WM3064 was provided by Anne Lanois and Alain Givaudan (Univ. Montpellier). TY was supported in part by the Government Scholarship to Study Abroad (GSSA), Taiwan. Computation was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. Tn5 and Phusion enzymes were procured through Protein Production Sweden. This project was supported by NGI OpenLab, Sweden. Funding for this work has been provided by a Wallenberg Academy Fellows grant to DL (https://kaw.wallenberg.org/en, 2021.0102).

Competing Interests

The authors declare no competing interests.

References

1. Northen TR, Kleiner M, Torres M, Kovács ÁT, Nicolaisen MH, Krzyżanowska DM, et al. Community standards and future opportunities for synthetic communities in plant-microbiota research. Nat Microbiol. 2024;9: 2774–2784. doi:10.1038/s41564-024-01833-4

2. Glandorf DCM, Brand I, Bakker PAHM, Schippers B. Stability of rifampicin resistance as a marker for root colonization studies of Pseudomonas putida in the field. Plant Soil. 1992;147: 135–142. doi:10.1007/bf00009379

3. Jansson JK. Tracking genetically engineered microorganisms in nature. Curr Opin Biotechnol. 1995;6: 275–283. doi:10.1016/0958-1669(95)80048-4

4. Rawls JF, Mahowald MA, Ley RE, Gordon JI. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell. 2006;127: 423–433. doi:10.1016/j.cell.2006.08.043

5. Schlechter RO, Kear EJ, Remus DM, Remus-Emsermann MNP. Fluorescent protein expression as a proxy for bacterial fitness in a high-throughput assay. Appl Environ Microbiol. 2021;87: e0098221. doi:10.1128/AEM.00982-21

6. Ordon J, Thouin J, Nakano RT, Ma K-W, Zhang P, Huettel B, et al. Chromosomal barcodes for simultaneous tracking of near-isogenic bacterial strains in plant microbiota. Nat Microbiol. 2024;9: 1117–1129. doi:10.1038/s41564-024-01619-8

7. Daniel BBJ, Steiger Y, Sintsova A, Field CM, Nguyen BD, Schubert C, et al. Assessing microbiome population dynamics using wild-type isogenic standardized hybrid (WISH)-tags. Nat Microbiol. 2024;9: 1103–1116. doi:10.1038/s41564-024-01634-9

8. Vo PLH, Ronda C, Klompe SE, Chen EE, Acree C, Wang HH, et al. CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. Nat Biotechnol. 2021;39: 480–489. doi:10.1038/s41587-020-00745-y

9. Shalev O, Karasov TL, Lundberg DS, Ashkenazy H, Pramoj Na Ayutthaya P, Weigel D. Commensal Pseudomonas strains facilitate protective response against pathogens in the host plant. Nat Ecol Evol. 2022. doi:10.1038/s41559-022-01673-7

10. Choi K-H, Gaynor JB, White KG, Lopez C, Bosio CM, Karkhoff-Schweizer RR, et al. A Tn7-based broad-range bacterial cloning and expression system. Nat Methods. 2005;2: 443–448. doi:10.1038/nmeth765

11. Rubin BE, Diamond S, Cress BF, Crits-Christoph A, Lou YC, Borges AL, et al. Species- and site-specific genome editing in complex bacterial communities. Nat Microbiol. 2022;7: 34–47. doi:10.1038/s41564-021-01014-7

12. Lundberg DS, de Pedro Jové R, Pramoj Na Ayutthaya P, Karasov TL, Shalev O, Poersch K, et al. Contrasting patterns of microbial dominance in the *Arabidopsis thaliana* phyllosphere. Proc Natl Acad Sci U S A. 2022;119: e2211881119. doi:10.1073/pnas.2211881119

13. Matsumoto H, Fan X, Wang Y, Kusstatscher P, Duan J, Wu S, et al. Bacterial seed endophyte shapes disease resistance in rice. Nat Plants. 2021;7: 60–72. doi:10.1038/s41477-020-00826-5

14. Basta T, Buerger S, Stolz A. Structural and replicative diversity of large plasmids from sphingomonads that degrade polycyclic aromatic compounds and xenobiotics. Microbiology. 2005;151: 2025–2037. doi:10.1099/mic.0.27965-0

15. Kaczmarczyk A, Vorholt JA, Francez-Charlot A. Synthetic vanillate-regulated promoter for graded gene expression in Sphingomonas. Sci Rep. 2014;4: 6453. doi:10.1038/srep06453

16. Bertram R, Hillen W. The application of Tet repressor in prokaryotic gene regulation and expression: Tet regulation in bacteria. Microb Biotechnol. 2008;1: 2–16. doi:10.1111/j.1751-7915.2007.00001.x

17. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. Practical innovations for high-throughput amplicon sequencing. Nat Methods. 2013;10: 999–1002. doi:10.1038/nmeth.2634

18. Kaczmarczyk A, Vorholt JA, Francez-Charlot A. Markerless gene deletion system for sphingomonads. Appl Environ Microbiol. 2012;78: 3774–3777. doi:10.1128/AEM.07347-11

19. Zhu L, Wu X, Li O, Qian C, Gao H. Cloning and characterization of genes involved in nostoxanthin biosynthesis of Sphingomonas elodea ATCC 31461. PLoS One. 2012;7: e35099. doi:10.1371/journal.pone.0035099

20. Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes: Gene dosage effects and genome organisation. Mol Microbiol. 2006;59: 1506–1518. doi:10.1111/j.1365-2958.2006.05046.x

21. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132. doi:10.1186/s13059-016-0997-x

22. Bernhards CB, Liem AT, Berk KL, Roth PA, Gibbons HS, Lux MW. Putative phenotypically neutral genomic insertion points in prokaryotes. ACS Synth Biol. 2022;11: 1681–1685. doi:10.1021/acssynbio.1c00531

23. Lambertsen L, Sternberg C, Molin S. Mini-Tn7 transposons for site-specific tagging of bacteria with fluorescent proteins. Environ Microbiol. 2004;6: 726–732. doi:10.1111/j.1462-2920.2004.00605.x

24. Barnes RJ, Leung KT, Schraft H, Ulanova M. Chromosomal gfp labelling of Pseudomonas aeruginosa using a mini-Tn7 transposon: application for studies of bacteria-host interactions. Can J Microbiol. 2008;54: 48–57. doi:10.1139/w07-118

25. Matsumoto A, Schlüter T, Melkonian K, Takeda A, Nakagami H, Mine A. A versatile Tn7 transposon-based bioluminescence tagging tool for quantitative and spatial detection of bacteria in plants. Plant Commun. 2022;3: 100227. doi:10.1016/j.xplc.2021.100227

26. Stern DL. Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites. bioRxiv. bioRxiv; 2016. p. 037762. doi:10.1101/037762

27. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26: 2460–2461. doi:10.1093/bioinformatics/btq461

28. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97: 6640–6645. doi:10.1073/pnas.120163297

29. Kuduvalli PN, Rao JE, Craig NL. Target DNA structure plays a critical role in Tn7 transposition. EMBO J. 2001;20: 924–932. doi:10.1093/emboj/20.4.924

30. Roos K, Werner E, Loessner H. Multicopy integration of mini-Tn7 transposons into selected chromosomal sites of a Salmonella vaccine strain: Multicopy integration of mini-Tn7 transposons. Microb Biotechnol. 2015;8: 177–187. doi:10.1111/1751-7915.12187

31. Jorrin B, Haskett TL, Knights HE, Martyn A, Underwood TJ, Dolliver J, et al. Stable, fluorescent markers for tracking synthetic communities and assembly dynamics. Microbiome. 2024;12: 81. doi:10.1186/s40168-024-01792-2

32. Silverman JD, Bloom RJ, Jiang S, Durand HK, Dallow E, Mukherjee S, et al. Measuring and mitigating PCR bias in microbiota datasets. PLoS Comput Biol. 2021;17: e1009113. doi:10.1371/journal.pcbi.1009113

33. Innerebner G, Knief C, Vorholt JA. Protection of Arabidopsis thaliana against leaf-pathogenic Pseudomonas syringae by Sphingomonas strains in a controlled model system. Appl Environ Microbiol. 2011;77: 3202–3210. doi:10.1128/AEM.00133-11

34. Song LCT, Alker ATP, Oromí-Bosch A, Swartz SE, Martinson JNV, Arora J, et al. Identification of proteins influencing CRISPR-associated transposases for enhanced genome editing. Sci Adv. 2026;12: eaea1429. doi:10.1126/sciadv.aea1429

35. Wickham H. Ggplot2: Elegant graphics for data analysis. New York, NY: Springer; 2016. Available: https://ggplot2.tidyverse.org

36. Addgene: Handling Plasmids from Addgene - Purifying Plasmid DNA. [cited 20 Sep 2024]. Available: https://www.addgene.org/protocols/purify-plasmid-dna/

37. Lundberg DS, Pramoj Na Ayutthaya P, Strauß A, Shirsekar G, Lo W-S, Lahaye T, et al. Host-associated microbe PCR (hamPCR) enables convenient measurement of both microbial load and community composition. Elife. 2021;10. doi:10.7554/eLife.66186

38. Gelsinger DR, Vo PLH, Klompe SE, Ronda C, Wang H, Sternberg SH. Bacterial genome engineering using CRISPR RNA-guided transposases. bioRxiv. 2023. doi:10.1101/2023.03.18.533263

39. Choi K-H, Kumar A, Schweizer HP. A 10-min method for preparation of highly electrocompetent Pseudomonas aeruginosa cells: application for DNA fragment transfer between chromosomes and plasmid transformation. J Microbiol Methods. 2006;64: 391–397. doi:10.1016/j.mimet.2005.06.001

40. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24: 2033–2040. doi:10.1101/gr.177881.114

41. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. Microb Genom. 2021;7. doi:10.1099/mgen.0.000685

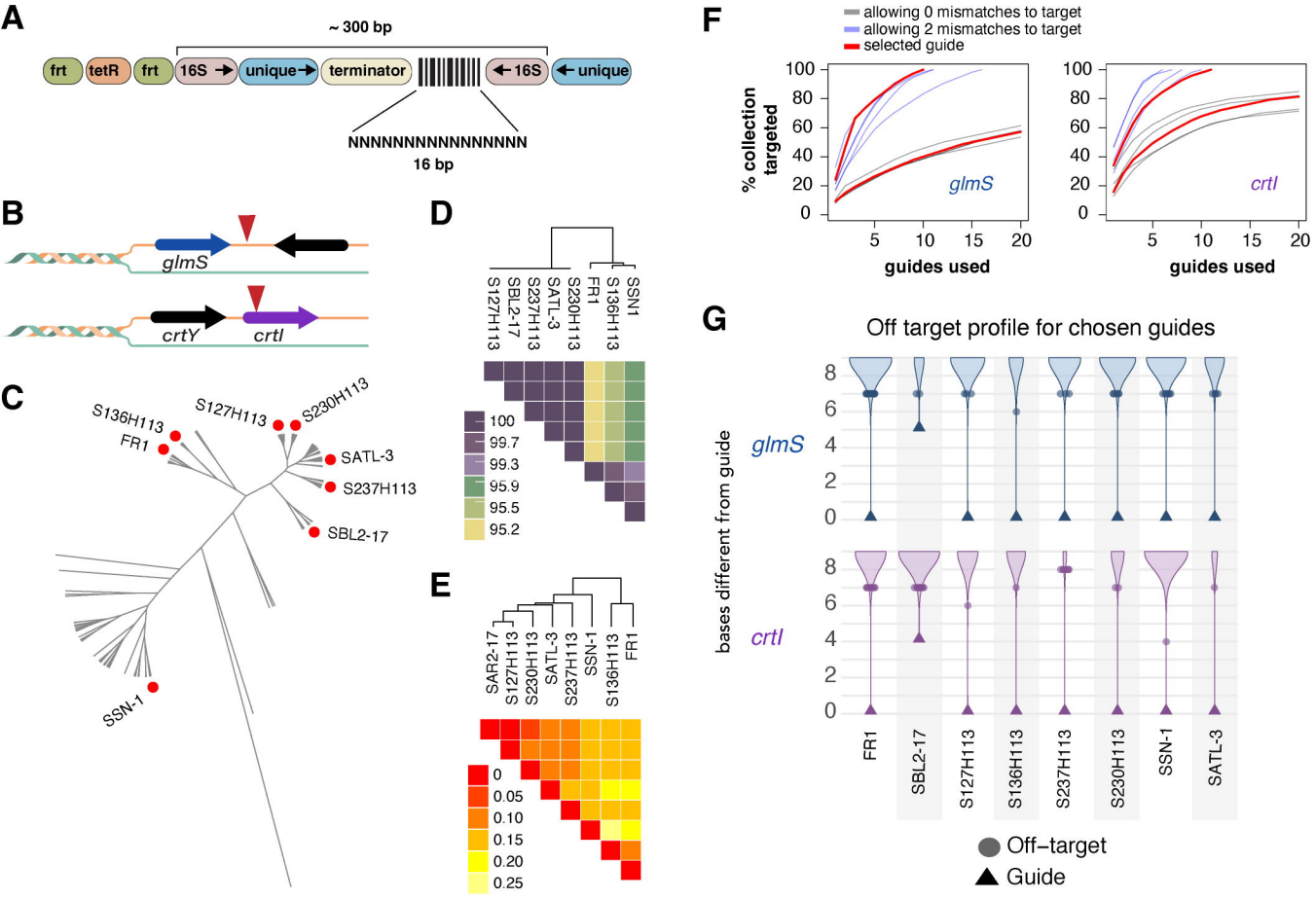
42. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4: e2584. doi:10.7717/peerj.2584

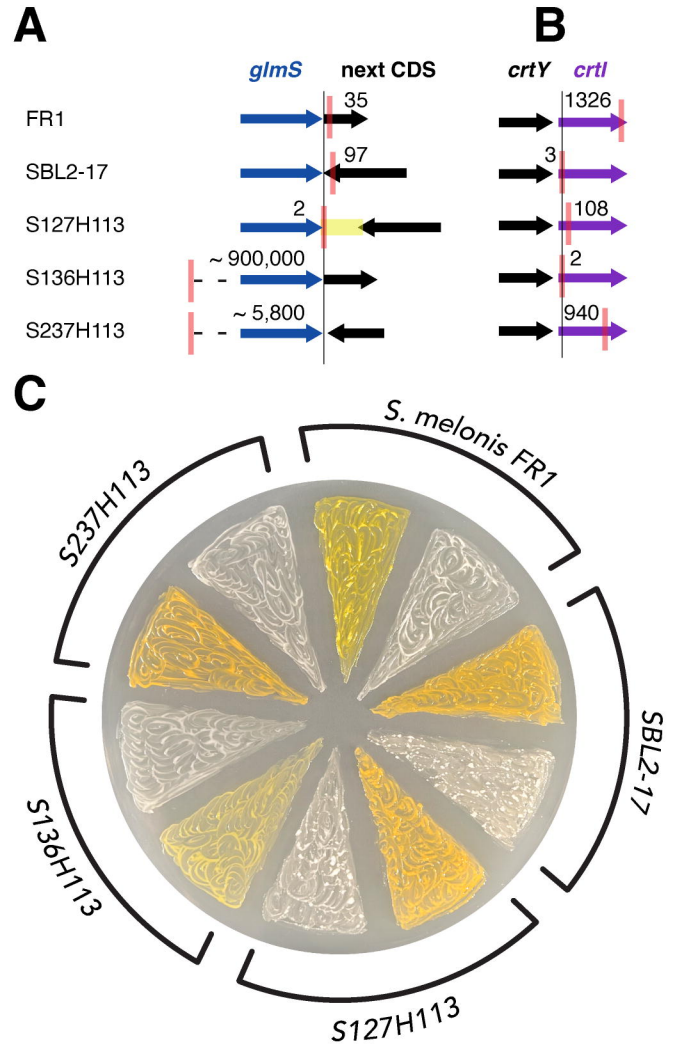
43. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. bioRxiv; 2016. p. 081257. doi:10.1101/081257

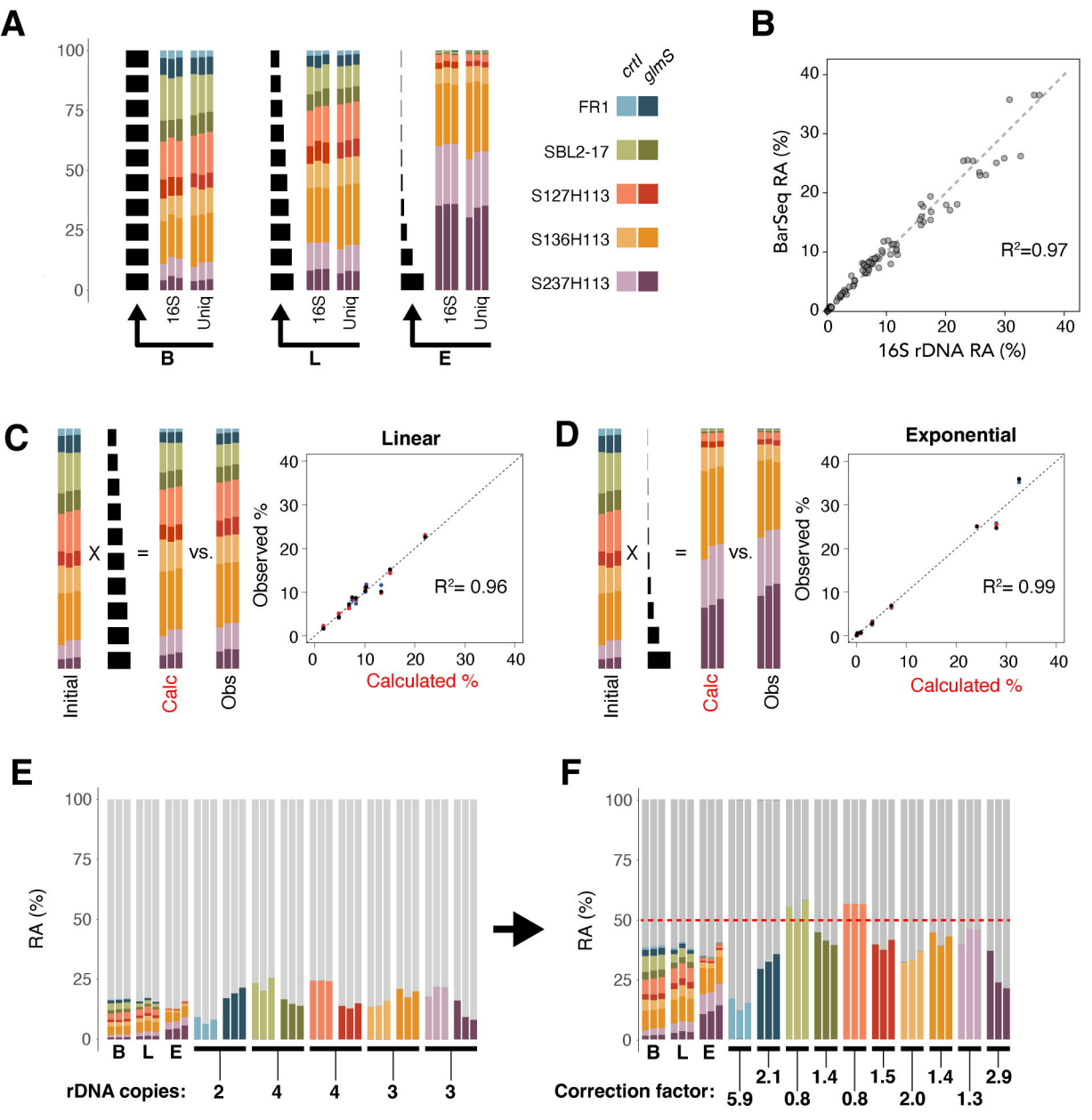
44. Edgar RC. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv. bioRxiv; 2016. p. 074161. doi:10.1101/074161

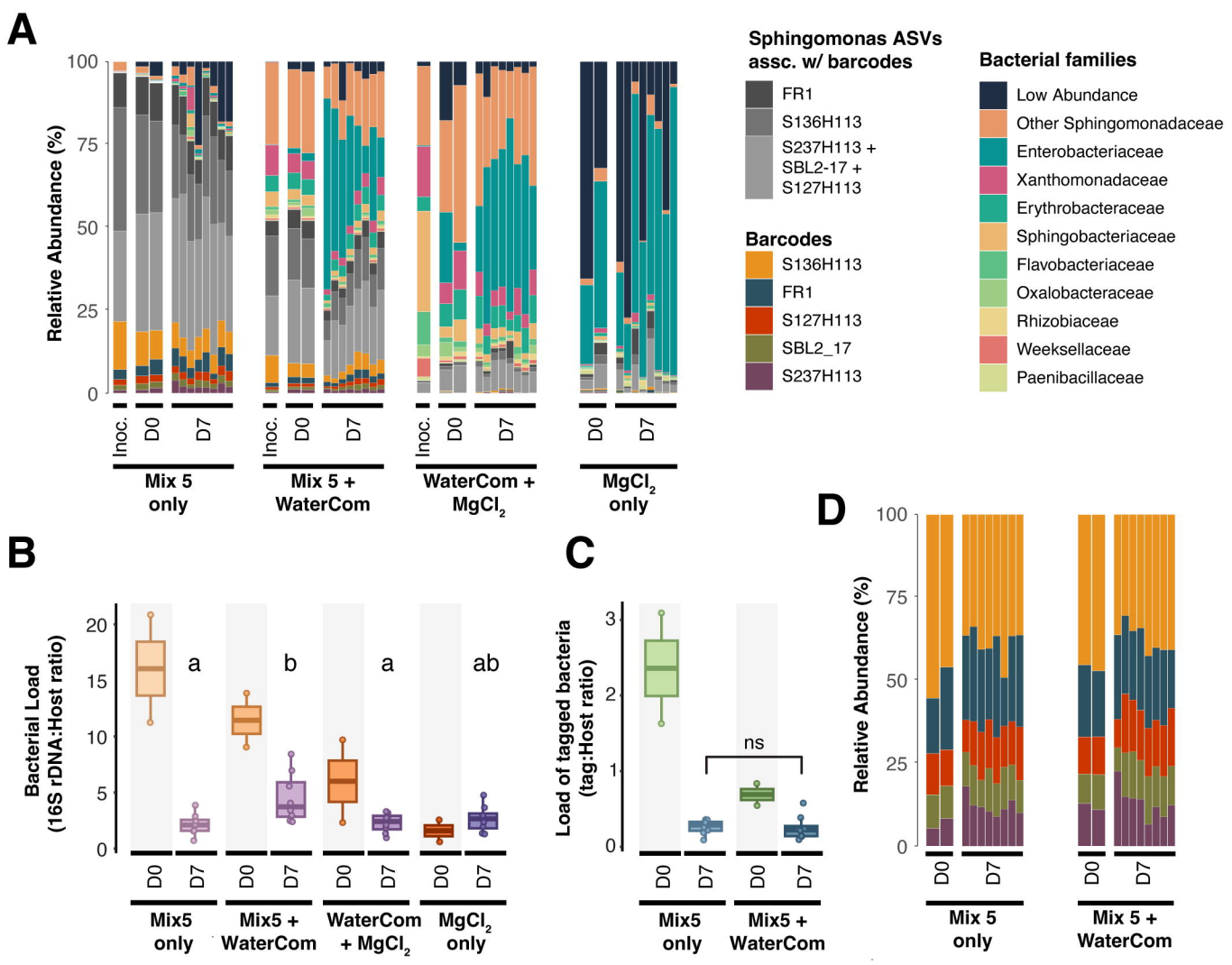
45. Wang Q, Cole JR. Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. Microbiol Resour Announc. 2024;13: e0106323. doi:10.1128/mra.01063-23

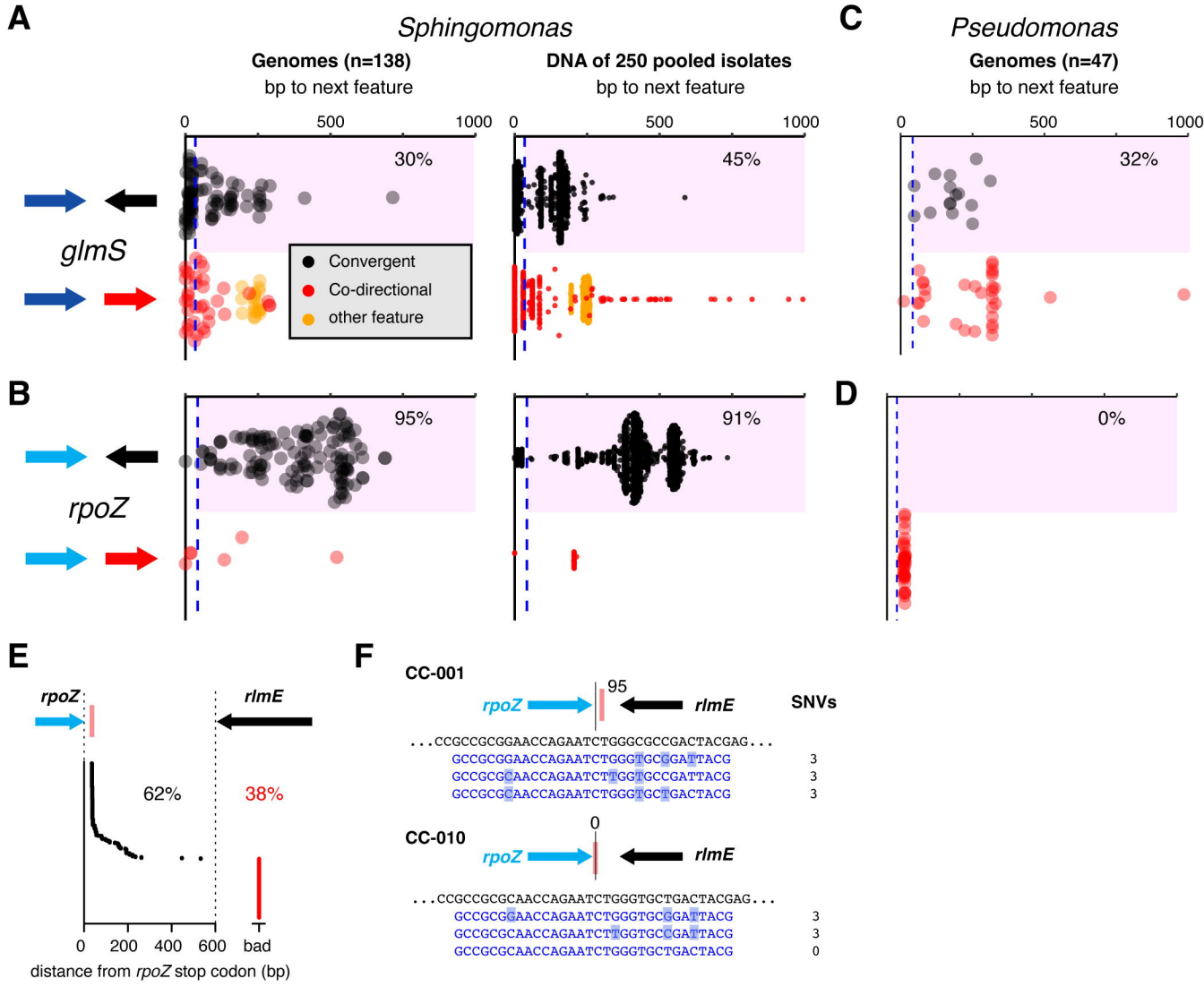
46. Wickham H. Reshaping Data with thereshapePackage. J Stat Softw. 2007;21: 1–20. doi:10.18637/jss.v021.i12


A
~ 300 bp
frt
tetR
frt
16S
unique
terminator
16S
unique
NNNNNNNNNNNNNNNN
16 bp
B
glmS
crtY
crtI
C
S136H113
FR1
S127H113
S230H113
SATL-3
S237H113
SBL2-17
SSN-1
D
S127H113
SBL2-17
S237H113
SATL-3
S230H113
FR1
S136H113
SSN1
100
99.7
99.3
95.9
95.5
95.2
E
SAR2-17
S127H113
S230H113
SATL-3
S237H113
SSN-1
S136H113
FR1
0
0.05
0.10
0.15
0.20
0.25
F
allowing 0 mismatches to target
allowing 2 mismatches to target
selected guide
% collection targeted
100
80
60
40
20
0
5
10
15
20
guides used
glmS
crtI
G
Off target profile for chosen guides
bases different from guide
glmS
crtI
8
6
4
2
0
FR1
SBL2-17
S127H113
S136H113
S237H113
S230H113
SSN-1
SATL-3
Off–target
Guide


A
B
glmS
next CDS
crtY
crtI
FR1
SBL2-17
S127H113
S136H113
S237H113
35
97
2
~ 900,000
~ 5,800
1326
3
108
2
940
C
S. melonis FR1
SBL2-17
S127H113
S136H113
S237H113


A
100
75
50
25
0
16S
Uniq
B
L
E
crtl
glmS
FR1
SBL2-17
S127H113
S136H113
S237H113
B
BarSeq RA (%)
16S rDNA RA (%)
R²=0.97
C
Initial
Calc
Obs
X
=
vs.
Linear
Observed %
Calculated %
R²= 0.96
D
Exponential
R²= 0.99
E
RA (%)
rDNA copies: 2 4 4 3 3
F
Correction factor: 5.9 2.1 0.8 1.4 0.8 1.5 2.0 1.4 1.3 2.9


A
Relative Abundance (%)
100
75
50
25
0
Inoc.
D0
D7
Mix 5 only
Mix 5 + WaterCom
WaterCom + MgCl2
MgCl2 only
Sphingomonas ASVs assc. w/ barcodes
FR1
S136H113
S237H113 + SBL2-17 + S127H113
Barcodes
S136H113
FR1
S127H113
SBL2_17
S237H113
Bacterial families
Low Abundance
Other Sphingomonadaceae
Enterobacteriaceae
Xanthomonadaceae
Erythrobacteraceae
Sphingobacteriaceae
Flavobacteriaceae
Oxalobacteraceae
Rhizobiaceae
Weeksellaceae
Paenibacillaceae
B
Bacterial Load (16S rDNA:Host ratio)
20
15
10
5
0
a
b
a
ab
D0
D7
Mix5 only
Mix5 + WaterCom
WaterCom + MgCl2
MgCl2 only
C
Load of tagged bacteria (tag:Host ratio)
3
2
1
0
ns
D0
D7
Mix5 only
Mix5 + WaterCom
D
Relative Abundance (%)
100
75
50
25
0
D0
D7
Mix 5 only
Mix 5 + WaterCom


A
Sphingomonas
Genomes (n=138)
bp to next feature
DNA of 250 pooled isolates
bp to next feature
C
Pseudomonas
Genomes (n=47)
bp to next feature
0
500
1000
30%
45%
32%
glmS
Convergent
Co-directional
other feature
B
95%
91%
0%
rpoZ
D
E
rpoZ
rlmE
62%
38%
0
200
400
600
bad
distance from rpoZ stop codon (bp)
F
CC-001
95
rpoZ
rlmE
SNVs
...CCGCCGCGGAACCAGAATCTGGGCGCCGACTACGAG...
GCCGCGGAACCAGAATCTGGGTGCGGATTACG 3
GCCGCGCAACCAGAATCTTGGTGCCGATTACG 3
GCCGCGCAACCAGAATCTGGGTGCTGACTACG 3
CC-010
0
rpoZ
rlmE
...CCGCCGCGCAACCAGAATCTGGGTGCTGACTACGAG...
GCCGCGGAACCAGAATCTGGGTGCGGATTACG 3
GCCGCGCAACCAGAATCTTGGTGCCGATTACG 3
GCCGCGCAACCAGAATCTGGGTGCTGACTACG 0