Species- and site-specific genome editing in complex bacterial communities

Benjamin E. Rubin[1,2,14], Spencer Diamond[1,3,14], Brady F. Cress[1,2,14], Alexander Crits-Christoph[4], Yue Clare Lou[1,4], Adair L. Borges[1,7], Haridha Shivram[1,2], Christine He[1,2,3], Michael Xu[1,2], Zeyi Zhou[1,2], Sara J. Smith[1,2], Rachel Rovinsky[1,2], Dylan C. Smock[1,2], Kimberly Tang[1,2], Trenton K. Owens[5], Netravathi Krishnappa[1], Rohan Sachdeva[1,3], Rodolphe Barrangou[6], Adam M. Deutschbauer[4,5], Jillian F. Banfield[1,3,7,8]* & Jennifer A. Doudna[1,2,9-13]*

[1]Innovative Genomics Institute, University of California, Berkeley, CA, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. [3]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA. [4]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA. [5]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [6]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA [7]Environmental Science, Policy and Management, University of California, Berkeley, CA, USA. [8]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia. [9]California Institute for Quantitative Biosciences, University of California, Berkeley, CA, USA. [10]Department of Chemistry, University of California, Berkeley, CA, USA. [11]Howard Hughes Medical Institute, University of California, Berkeley, CA, USA. [12]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA [13]Gladstone Institutes, University of California, San Francisco, CA, USA. [14]These authors contributed equally to this work: Benjamin E. Rubin, Spencer Diamond, Brady F. Cress.

*e-mail: doudna@berkeley.edu; jbanfield@berkeley.edu

Knowledge of microbial gene functions comes from manipulating the DNA of individual strains in isolation from their natural communities. While this approach to microbial genetics has been foundational, its requirement for culturable microorganisms has left the majority of microbes and their interactions genetically unexplored. Here, we describe a generalizable strategy for editing the genomes of specific organisms within microbial communities. We identified genetically tractable bacteria within a community using Environmental Transformation Sequencing (ET-Seq), an approach in which non-targeted transposon integrations are mapped and quantified following community delivery. We next developed and used DNA-editing All-in-one RNA-guided CRISPR-Cas Transposase (DART) systems for targeted DNA insertion into organisms identified as tractable by ET-Seq, enabling organism- and locus-specific genetic manipulation within the community context. To illustrate the utility of our approach, we selectively edited closely related strains, measured gene fitness, and enriched targeted members within soil and infant gut microbiota. These results establish a new paradigm for targeted community editing relevant to research and applications on medical, agricultural, and industrial microbiomes.

Genetic mutation and observation of phenotypic outcomes are the primary means of deciphering gene function in microorganisms. This classical genetic approach requires manipulation of isolated species, limiting knowledge in three fundamental ways. First, the vast majority of microorganisms have not been isolated in the laboratory and are thus largely untouched by molecular genetics[1]. Second, genes involved in interactions between microorganisms remain mostly unexplored[2]. Third, microorganisms grown and studied in isolation quickly adapt to their new lab environment, obscuring their true “wild type” physiology[3]. Since most microorganisms relevant to the environment, industry, and health live in communities, approaches for precision genome editing in community contexts will be transformative.

Advances toward genome editing within microbial communities have included assessing gene transfer to microbiomes using selectable markers[4–9], microbiome manipulation leveraging pre-modified isolates or exogenously introduced targets[10,11], and use of temperate phage for species-specific integration of genetic payloads[12,13]. However, a generalizable strategy for programmable organism- and locus-specific editing within a community of wild-type microbes has not yet been reported[14].

Here, we show that individual organisms within microbial communities can be targeted for site-specific genome editing, enabling manipulation of species without requiring prior isolation or engineering. Using a method developed for this study, Environmental Transformation Sequencing (ET-Seq), we identified genetically accessible species within a synthetic soil community assembled from isolates without the application of selection. These results enabled targeted genome editing of microbes in this consortium using DNA-editing All-in-one RNA-guided CRISPR-Cas Transposase (DART) systems developed here. We then applied these tools to track the fitness of a genetic mutant created inside the soil community without selectable markers. Furthermore, we demonstrated targeted editing with selectable markers in both the synthetic soil community and an infant gut microbiota allowing for subsequent enrichment and isolation of the edited members. The species-specific editing described here lays the foundation for both experimentation and control of organisms within their native communities.

ET-Seq identifies genetically accessible microbial community members

Editing organisms within a complex microbiome requires knowing which constituents are accessible to nucleic acid delivery and editing. We developed ET-Seq to assess the ability of individual species within a microbial community to acquire and integrate exogenous DNA (Fig. 1a). In ET-Seq, a microbial community is exposed to a randomly integrating mobile genetic element (here, a *mariner* transposon), and in the absence of any selection, total community DNA is extracted and sequenced using two protocols. In the first, we enrich and sequence the junctions between the inserted and host DNA to determine insertion location and quantity in each host (Methods). In the second protocol, we conduct low-depth metagenomic sequencing to quantify the abundance of each community member in a sample (Extended Data Fig. 1a). If the community has not been previously sequenced, high-depth metagenomic sequencing would be required at this step to provide reference genomes as well as abundance information. Together, these sequencing procedures provide relative insertion efficiencies for microbiota members. To convert this relative measurement into one anchored to a known insertion efficiency we normalize these data according to an internal standard which we add in a uniform amount to every sample. The standard consists of DNA from a transposon mutant library that was generated with antibiotic selection and thus contains an insertion in every genome. The final output of ET-Seq estimates the proportion of each organism's population that harbored transposon insertions at the time DNA was extracted, a combined measure of delivery, insertion efficiency, and mutant survival within the delivery condition (Extended Data Fig. 1b). To facilitate the analysis of these disparate data, we developed a complete bioinformatic pipeline for quantifying insertions and normalizing results according to both the internal control and metagenomic abundance (https://github.com/SDmetagenomics/ETsuite and Methods). Together, the experimental and bioinformatic approaches of ET-Seq reveal species-specific genetic accessibility by measuring the percentage of each member of a given microbiome that acquires a transposon insertion.

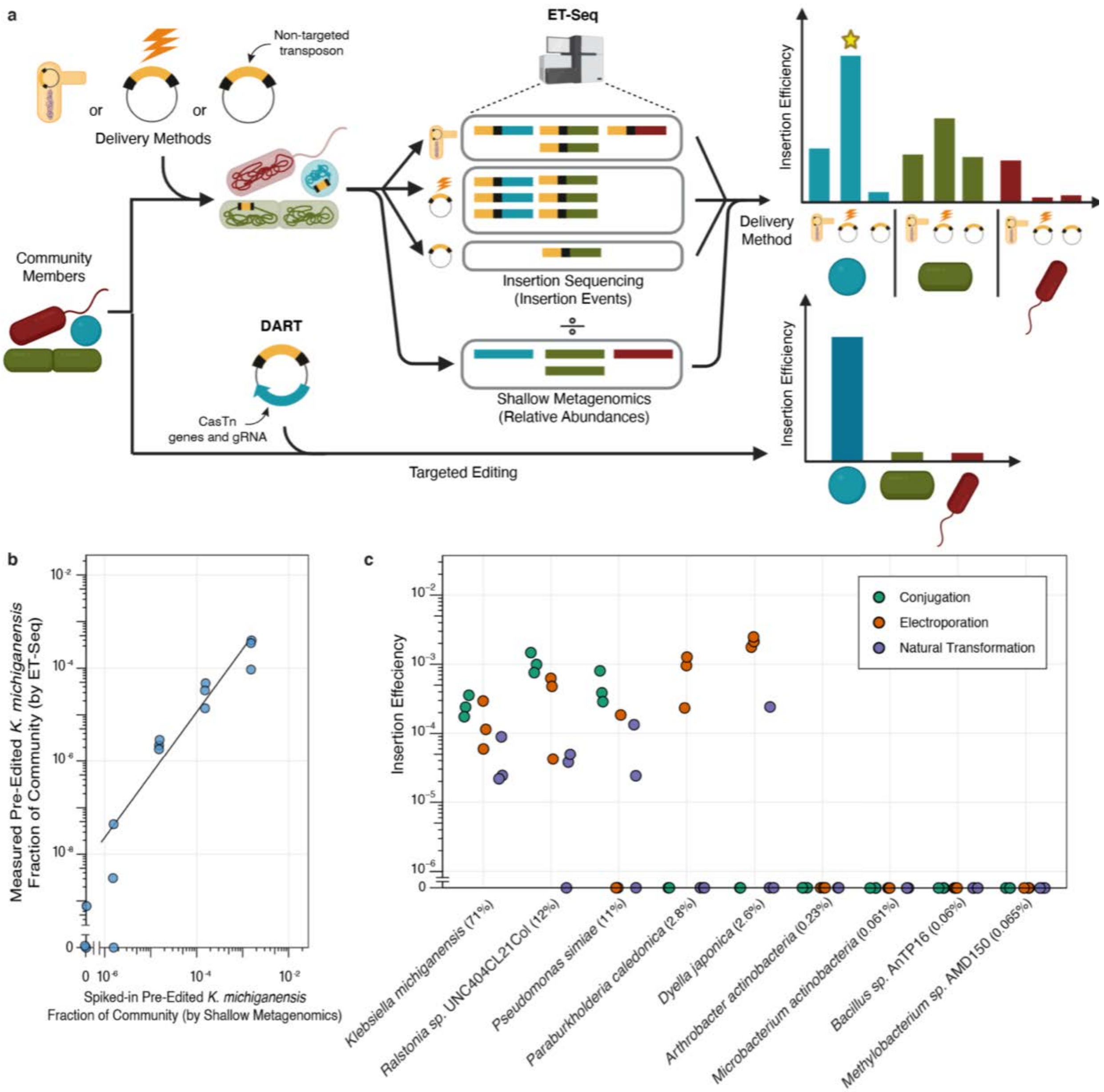


Fig. 1 | ET-Seq for quantitative measurement of insertion efficiency in a microbial community. a, ET-Seq provides data on insertion efficiency of multiple delivery approaches, including conjugation, electroporation, and natural DNA transformation, on microbial community members. In this illustrative example, the blue strain is most amenable to electroporation (star). This data allows for the determination of feasible targets and delivery methods for DART targeted editing. **b,** ET-Seq determined efficiencies for known quantities of spiked-in pre-edited *K. michiganensis*. Solid line is the fit of the linear regression to the data not including zeros (n = 11 independent samples; R^2 = 0.89). **c,** ET-Seq determined insertion efficiencies (insertion containing portion within each species) for conjugation, electroporation, and natural transformation of the synthetic soil community (n = 3 biological replicates). Average relative abundance for each organism is indicated in parentheses.

ET-Seq was developed and tested on a nine-member microbial consortium made up of bacteria from three phyla that are often detected and play important metabolic roles within soil microbial communities (Supplementary Table 1). We initially tested the accuracy and detection limits of ET-Seq by adding to the synthetic soil community a known amount of a previously prepared *mariner* transposon library of one of its member species, *Klebsiella michiganensis* M5a1 (*K. michiganensis*). The ET-Seq-derived portion of cells receiving insertions closely correlated to the known fractions of edited *K. michiganensis* present in each sample (Fig. 1b; $R^2 = 0.89$). It is of note that in one of the replicates a low insertion frequency is detected at the no-spike in control. Such false positives appear to be the result of chimeras that can form during library preparation and we have taken extensive bioinformatic steps to filter out approximately 85% of these events (Methods and Extended Data Fig. 2). However, given the presence of low-level artifacts that escape filtering methods, consistent insertion detection across all replicates is an important indicator of transformation success and confidence. These data demonstrate quantification of genetic insertions that occur in cells making up 0.001% of the estimated total population, which is 10-100X more rare than those detected by typical rare-variant detection strategies[15].

We next used ET-Seq to compare insertion efficiencies in the synthetic soil community after *mariner* transposon delivery by conjugation, natural transformation with no induction of competence, or electroporation of the transposon vector. We measured insertions made by at least one delivery strategy reproducibly in the five species that grew to make up over 99% of the community (Fig. 1c, Extended Data Fig. 3, and Supplementary Table 2). Even for *Paraburkholderia caledonica* and *Dyella japonica* UNC79MFTsu3.2, which each make up ~2.5% of the community, we could measure insertions by electroporation. We detected no insertions in the remaining community members, which were likely below ET-Seq's limit of detection given these members' extreme rarity (<0.5%). We also identified preferred delivery methods to produce insertions in certain members. Electroporation mutants were reproducibly measured for *P. caledonica* and *D. japonica*, while mutants made by the other methods were not. In contrast, conjugation mutants were reproducibly measured for *Pseudomonas simiae* WCS417 (*P. simiae*), while other methods were not. These results show that ET-Seq can identify and quantify genetic manipulation of microbial community members and reveal suitable DNA delivery methods for each.

Targeted genome editing with CRISPR-Cas transposases

The ability to programmably introduce genome edits to a single type of organism in a microbial community and to target those edits to a defined location within its genome would be a foundational advance in microbiological research with many useful applications. We reasoned that RNA-guided CRISPR-Cas Tn7 transposases could provide the ability to both ablate function of targeted genes and deliver customized genetic cargo in organisms shown to be genetically tractable by ET-Seq[11,16–18] (Fig. 1a). However, the two-plasmid ShCasTn[16] and three-plasmid VcCasTn[17] systems are not amenable to efficient delivery within complex microbial communities or even beyond *E. coli* due to their multiple plasmids. Since ET-Seq identified conjugation and electroporation as broadly effective delivery approaches in the tested communities, we designed and constructed all-in-one conjugative versions of these CasTn vectors that could be used for delivery by either strategy (Fig. 2a and Methods). These DART systems are comparable to the INTEGRATE system[11], but are barcoded and compatible with the same sequencing methods used for ET-Seq. The barcodes, present on all transposons used in this paper, allow for the detection and tracking of uniquely edited cells. Thus, DART can be used seamlessly with ET-Seq to rapidly assay the efficacy of CRISPR-Cas-guided transposition into the genome of a target organism in the absence of selectable markers.

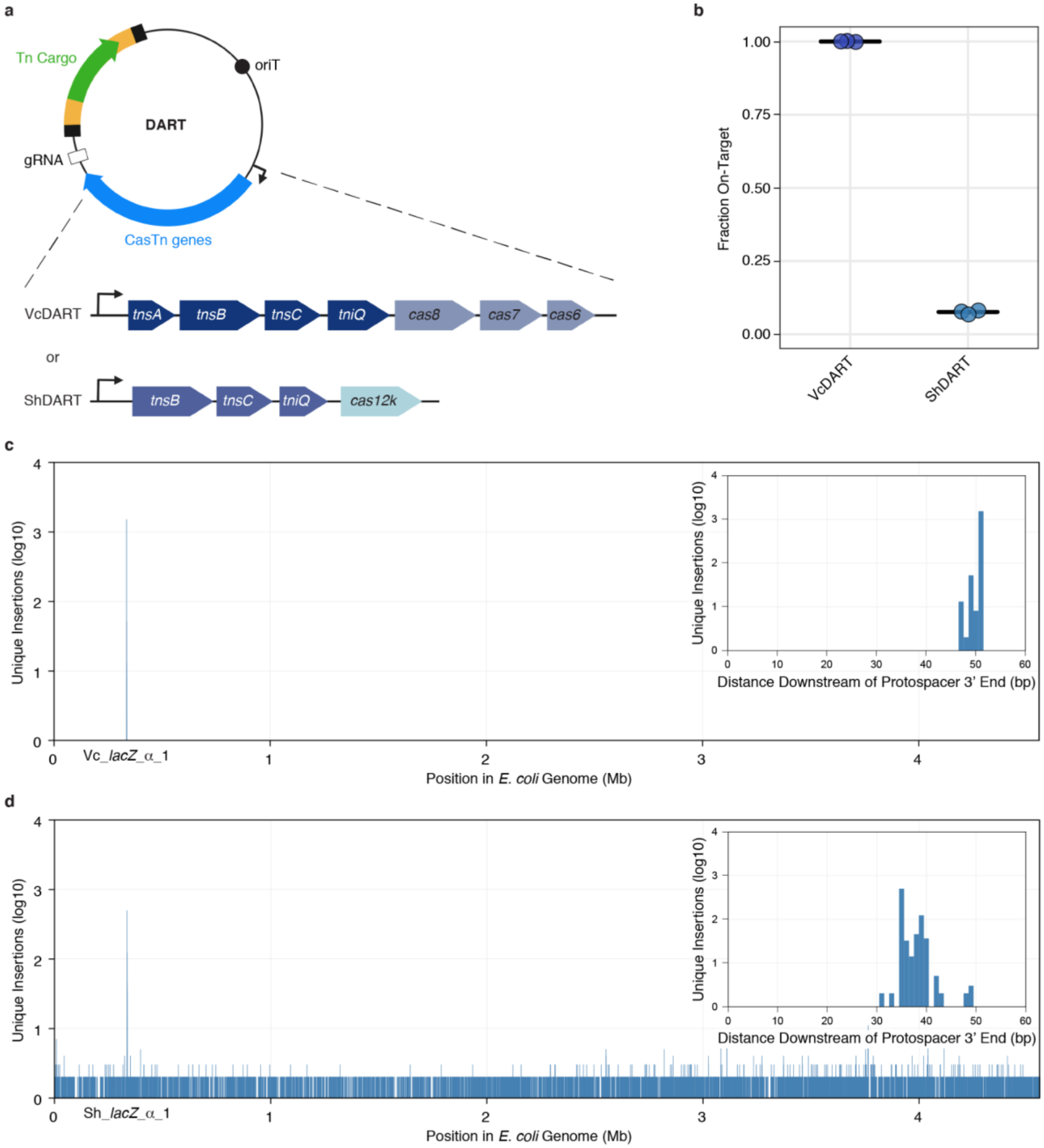


Fig. 2 | Benchmarking all-in-one conjugative targeted vectors. a, Schematic of VcDART and ShDART delivery vectors. **b,** Fraction of insertions that occur 200 bp downstream of the 3' end of the protospacer target site. Mean for three independent biological replicates is shown as cross bars. **c-d,** Aggregate unique insertion counts (n = 3 biological replicates) across the *E. coli* BL21(DE3) genome, determined by presence of unique barcodes, using **c,** VcDART and **d,** ShDART. The inset shows a 60 bp wind ow downstream of the target site where the peak of targeted insertions was observed. Insertion distance downstream of the target site is calculated from the 3' end of the protospacer.

We first compared the transposition efficiency and specificity of the DART systems in *E. coli* to identify the most promising candidate for targeted genome editing in microbial communities. VcDART and ShDART systems harboring Gm^R cargo with a *lacZ*-targeting or non-targeting guide RNA were conjugated into *E. coli* to quantify transposition efficiency (Methods and Extended Data Fig. 4a). Furthermore, we adapted ET-Seq's ability to measure non-targeted insertions within communities to characterize target site specificity of DART systems following outgrowth of transconjugants in selective medium. While ShDART yielded approximately tenfold more colonies possessing insertions than VcDART (Extended Data Fig. 4b), >92% of the selectable colonies obtained using ShDART were off-target, compared to no detectable off-target insertions for VcDART of the 1,586 independent insertions measured via ET-Seq (Fig. 2b-d). The lack of any detectable chimeras in these data is likely the result of the amplification of signal provided by outgrowth before measurement. Considering VcDART's high on-target specificity, it is notable that its insertion efficiency in *E. coli* is similar to the widely used non-targeted *mariner* transposon (Extended Data Fig. 4d). Further attempts to optimize the editing efficiency by substituting stronger DART promoters did not lead to increased transposition (Extended Data Fig. 4c-f). Additionally, all ShDART vectors with non-targeting guides led to similar levels of transposition as *lacZ*-targeting guides (Extended Data Fig. 4e-f), consistent with previous results showing non-targeted transposition produced by the ShCAST Tn7 system, even in absence of Cas12k[16]. Due to VcDART's high target site specificity and insertion efficiency, we focused on VcDART to test the potential for targeted microbial community genome editing.

Targeted species- and locus-specific community editing by programmable transposition

We reasoned that RNA-programmed transposition could be deployed for targeted editing of species within a microbial consortium. As an abundant member shown by ET-Seq to be tractable by conjugation (Fig. 1c), we first targeted *K. michiganensis*. Conjugation was used to introduce the VcDART vector into the community with multiplexed guide RNAs specific to two locations in the *K. michiganensis* genome (Methods). Following delivery of VcDART, and in absence of selection, ET-Seq detected insertions of the barcoded, marker-free transposon at the targeted loci (Fig. 3a; p_{Site1} = 1.30e-4; p_{Site2} = 1.33e-8; exact poisson probability). Loci outside of the targeted sites are likely representative of chimeras as they each contain only a single insertion in one replicate and are statistically insignificant (p = 0.285; exact poisson probability). This highly accurate community editing is further supported by later sequencing of selectable VcDART-edited colonies showing exclusively on-target mutations among 96 colonies sampled. Thus, targeted and programmable edits can be made, multiplexed, and detected without selection in a non-model species within a consortium.

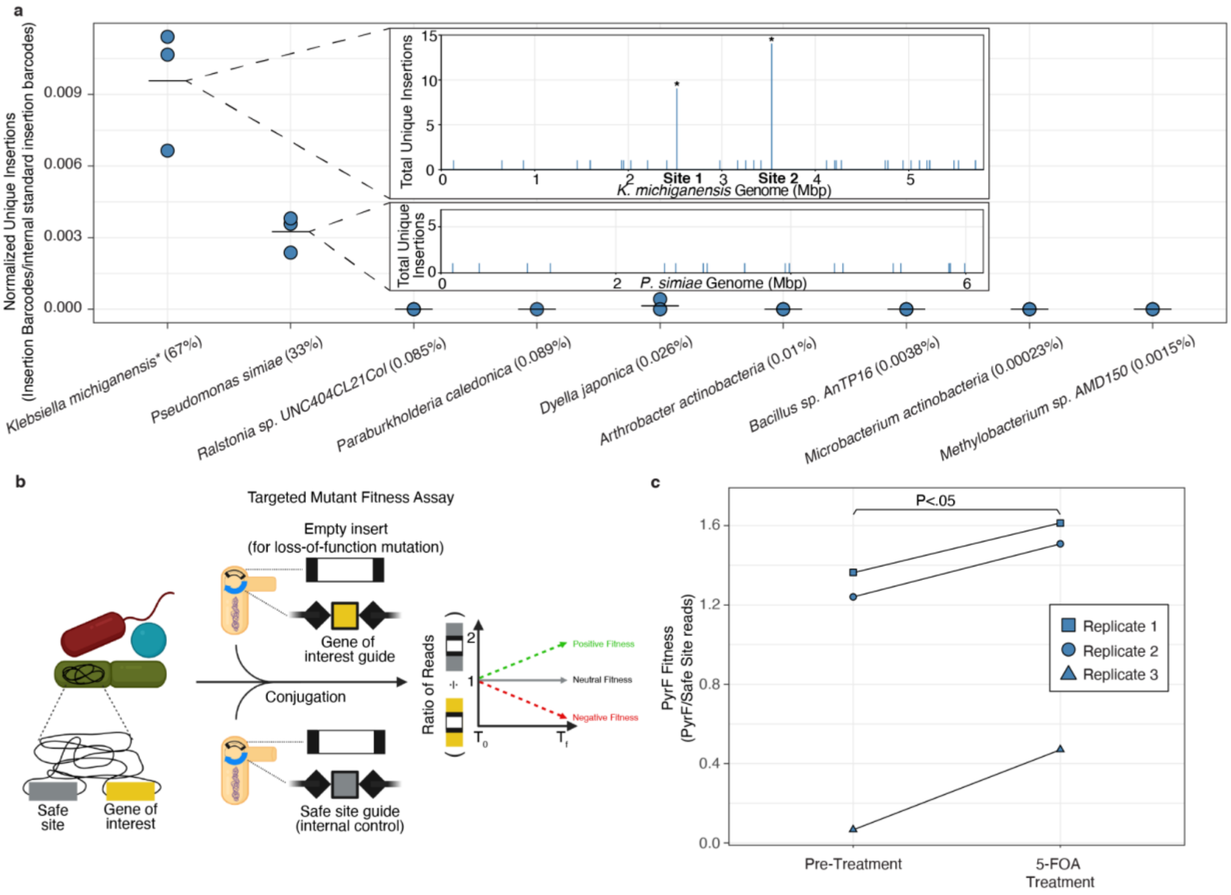


Fig. 3 | Selection free targeted editing and mutant tracking in the synthetic soil consortium. a, The main figure shows the number of insertions detected by ET-Seq in each species normalized for sequencing effort by the *B. thetaiotaomicron* internal standard. The insets show the location of unique insertions summed for the three replicates in *K. michiganensis* (upper) and *P. simiae* (lower) *p<0.001 Poisson Probability. **b,** The diagram shows the use of ET-Seq to quantify the fitness effect of a VcDART mutation of interest, measured as the ratio of mutant of interest reads normalized to Safe Site mutant reads at the assay end point divided by their ratio at the beginning. **c,** Fitness of *pyrF* mutant under 5-FOA treatment as measured by the ratio of *pyrF* to Safe Site reads. Lines connect biologically paired replicates sampled longitudinally.

Tagging and tracking mutant fitness in a microbial community

Strategies do not yet exist for performing traditional genetic assays of gene function within a community. We tested whether VcDART combined with ET-Seq enables targeted, selection-free tagging and tracking of genetic mutant fitness inside of a microbial community (Fig. 3b). Demonstration of this targeted genetic mutant fitness assay was conducted by using ET-Seq to track the relative fitness of the two edits made to *K. michiganensis* (Fig 3a). The mutations contain a barcoded VcDART transposon disrupting a gene of interest, *pyrF*, and an internal control locus, referred to here as a Safe Site due to its predicted fitness neutrality (Methods). The *pyrF* gene is commonly used as an endogenous counter-selectable marker. We predict that disruption of the targeted *pyrF* homolog in *K. michiganensis* will facilitate faster growth in the presence of growth inhibitory 5-fluoroorotic acid (5-FOA). The edited community was grown in presence of 5-FOA, and ET-Seq was used to quantify the effect of this condition on the fitness of the *pyrF* mutant in the community. As expected, ET-Seq detected higher fitness of *pyrF* mutants relative to Safe Site mutants in the presence of 5-FOA (Fig. 3c; p = 0.025; paired t-test). Thus, the combined efficiency of VcDART editing and ET-Seq detection sensitivity represents a powerful tool for probing gene fitness and function directly within microbial communities without perturbations caused by traditional mutant selection requirements.

Species-specific genetic tagging and isolation from a community

To demonstrate another application of our editing tools within a community, we used selectable edits to *K. michiganensis* and *P. simiae* to separately isolate each member from the community. Similar to the previous experiment, insertions were designed to produce loss-of-function mutations in the *K. michiganensis* and *P. simiae pyrF* genes (Fig. 4a). In this experiment, however, transposons carried two antibiotic resistance markers conferring resistance to streptomycin and spectinomycin (*aadA*) and carbenicillin (*bla*). Together, the simultaneous loss-of-function and gain-of-function mutations allowed for a strong selective regime. VcDART targeted to *K. michiganensis* or to *P. simiae pyrF* followed by selection led to enrichment of these organisms, each to >99% pure culture (Fig. 4b). No outgrowth was detected when using a guide RNA that did not target these respective microbial genomes. Recovered transformant colonies of *K. michiganensis* and *P. simiae* analyzed by PCR and Sanger sequencing showed full length, *pyrF*-disrupting VcDART transposon insertions 48-50 bp downstream of the guide RNA target site (Extended Data Fig. 5), which is consistent with previously characterized spacer-insert distances (Fig. 2c). These results demonstrate how targeted edits can be used to enrich and isolate specific bacteria from a community.

We next tested VcDART as a means to confer a new metabolic capability to a targeted organism in a microbial community. Whereas VcDART cargo containing antibiotic markers enables positive selection of edited cells while negatively selecting against unedited community members, cargo containing a nutrient utilization pathway can provide a metabolic niche for edited cells while minimally impacting untargeted members. To demonstrate targeted metabolic pathway integration, we used VcDART to provide *P. simiae* with lactose assimilation capacity inside a four-member community otherwise incapable of metabolizing lactose (Fig. 4c). Specifically, VcDART with a guide RNA targeting the *P. simiae* Safe Site and containing constitutive lactose permease (*lacY*) and beta-galactosidase (*lacZ*) cargo enabled growth of *P. simiae* on minimal medium containing lactose as a sole carbon source (Fig. 4d). It is of interest that *Ralstonia* sp. UNC404CL21Col (hereafter *Ralstonia* sp.), which could not grow alone on lactose medium, constitutes ~5% of the final culture indicating that cross-feeding by the edited *P. simiae* may allow *Ralstonia* sp. to grow. To test whether off-target insertions in *Ralstonia* sp. may alternatively explain its growth in lactose, we performed shotgun metagenomic sequencing on each of the three replicate samples. We identified only a single read pair in one replicate supporting an insertion junction in *Ralstonia* sp. despite high coverage of its genome (Total = 203x coverage; Avg = 68x coverage). This was significantly less (P-value = 0.00058; two-sample t-test) than the 385 read pairs supporting insertion junctions in *P. simiae* across the replicates suggesting off-targets are likely not a major source of *Ralstonia* sp. presence (Extended Data Fig. 6). Therefore, targeted addition of a metabolic niche can be used as an enrichment tool, which is likely to be applicable in more communities and with less disturbance to unedited members than antibiotic selection.

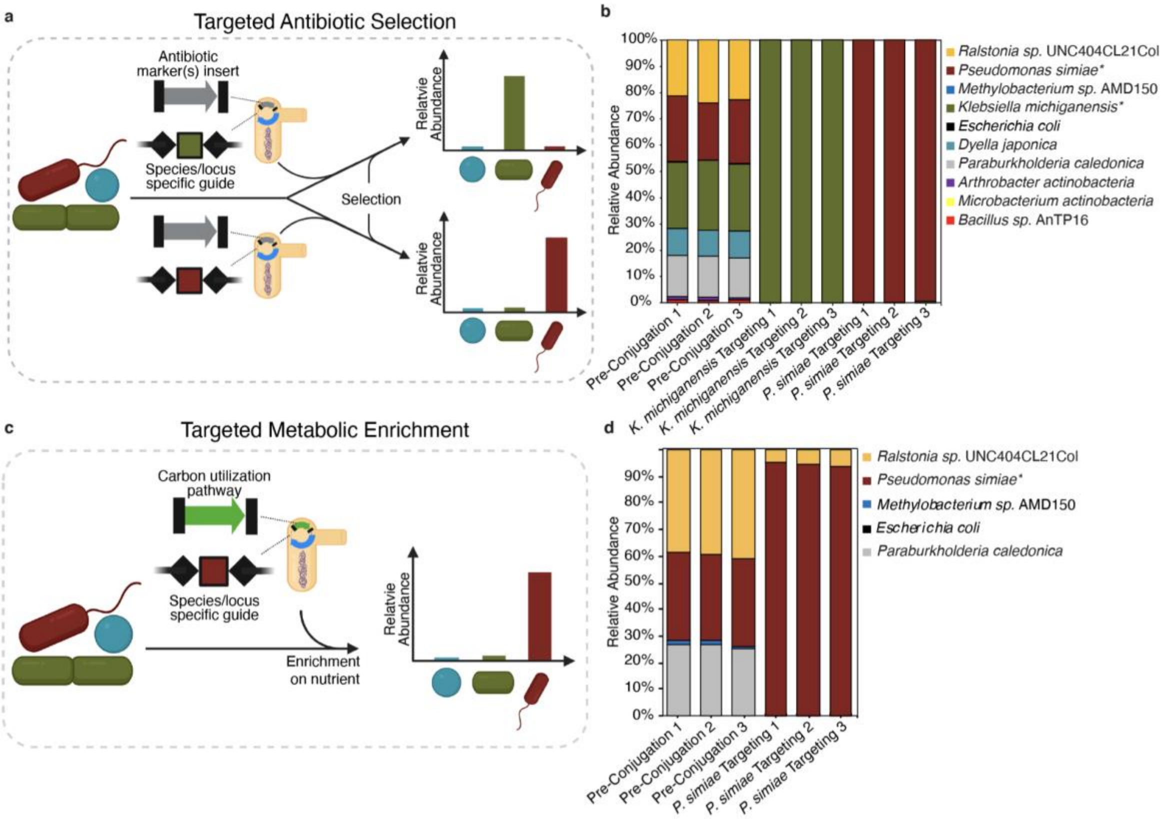


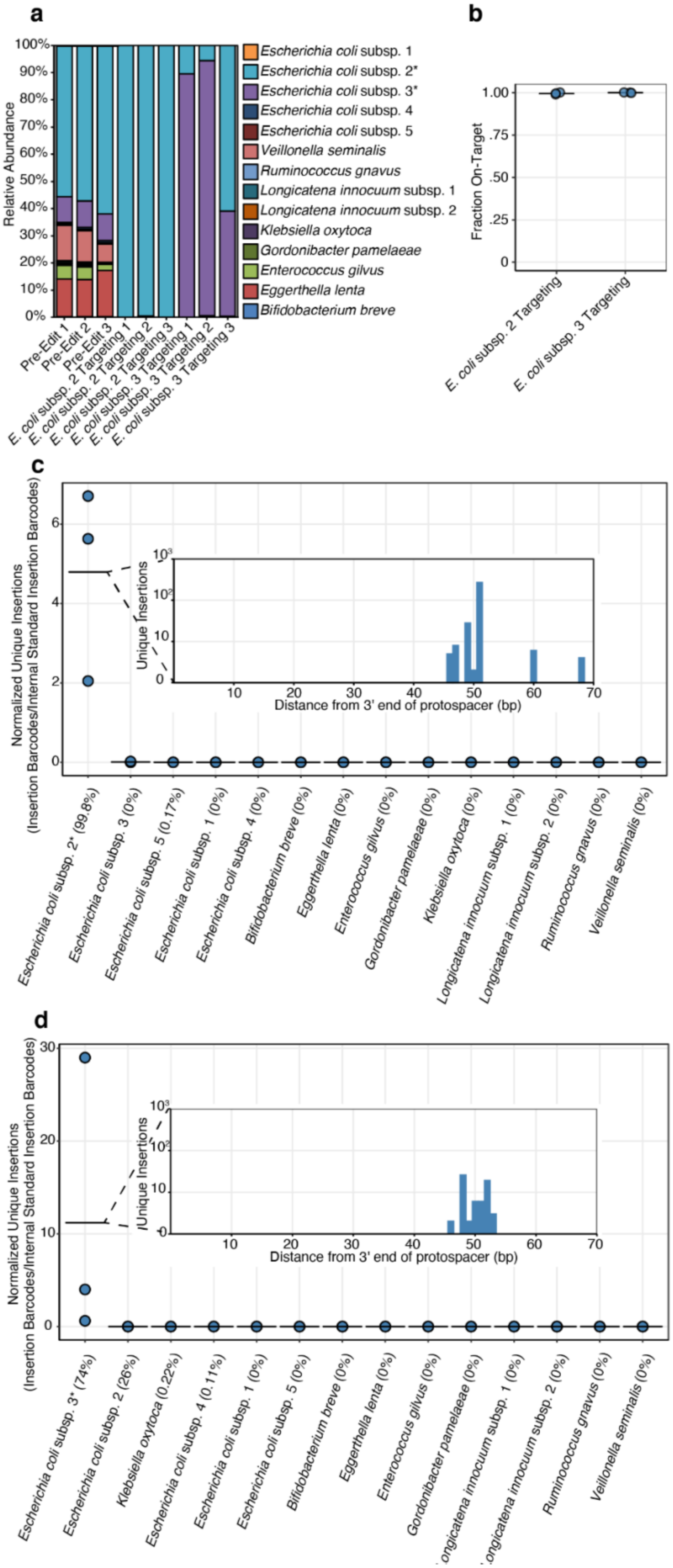
Fig. 4 | Enrichment of targeted strains in microbial communities. a, VcDART delivery of antibiotic markers into a microbial community using species-specific crRNA, followed by selection for transposon cargo, facilitates isolation of targeted organisms. **b,** Relative abundance of synthetic soil community constituents measured by metagenomic sequencing before conjugative VcDART delivery and after selection for *pyrF*-targeted antibiotic casette in *K. michiganensis* or *P. simiae*. **c,** VcDART delivery of a nutrient utilization pathway, guided by species-specific crRNA, into a microbial community facilitates enrichment of a targeted organism through growth on the appropriate nutrient. **d,** Relative abundance of the constituents of a four-member community incapable of utilizing lactose measured before conjugative VcDART delivery and after lactose-based enrichment for Safe Site-targeted lacZY transposition into *P. simiae*.

Strain-specific editing in an infant gut enrichment

To apply the community editing technologies developed here to a health relevant microbial community that has not been reduced to isolates, we developed a human infant gut microbiota. A stool sample from a 90-day-old infant, previously collected as part of a large scale metagenomic study[19] was used as inoculum. Mapping of metagenomic reads collected from the infant gut community to a reference set of 1005 genomes from that study identified 14 genomes represented above 0.1% relative abundance, representing all of the phyla (Firmicutes, Actinobacteria, and Proteobacteria) present in the original stool sample (Extended Data Fig. 7). Among the genomes detected are five strains of *E. coli* including members of the B2 and D phylogroups, the predominant groups of virulent extraintestinal *E. coli*[20] (Supplementary Table 1). ET-Seq of the community showed that *E. coli* was the only constituent receiving insertions (Extended Data Fig. 8). This result is unsurprising because the other members of the community above 1% relative abundance have not been shown to be editable in isolation.

Fig. 5 | Strain-resolved targeted editing in the infant gut microbiota. a, Relative abundance of the infant gut community before and after VcDART editing and selection for targeted loci within *E. coli* subsp. 2 and 3. **b,** Fraction of insertions that occur within 20 bp of the expected target site (50 bp downstream of the 3' end of protospacer). **c-d,** Unique insertion locations for targeted loci within **c,** *E. coli* subsp. 2 and **d,** *E. coli* subsp. 3. The main figures show unique insertions detected by ET-Seq normalized by the *B. thetaiotaomicron* internal standard. The insets show aggregate unique insertion counts (n = 3 biological replicates) within the protospacer adjacent region. In **a** and **c-d** members with relative abundance above

0.1% are shown and the targeted *E. coli* subsp. is noted with asterisks.



Using VcDART, we targeted unique loci within the community's *E. coli* strains that allowed for enrichment and isolation of specific strains carrying the loci of interest. We began by identifying clinically relevant sites that existed within the population of *E. coli* strains[21,22]. The loci, upstream of a fimbriae gene cluster and within a propanediol utilization gene cluster (Extended Data Fig. 9a), were targeted for editing with a selection marker conferring resistance to streptomycin and spectinomycin (*aadA*) and carbenicillin (*bla*). After selection, we performed metagenomic sequencing on the enrichments and were able to *de novo* assemble high quality genomes (*E. coli* subsp. 2 and 3) of the two strains containing the targeted loci (average of 99.93% completeness and 0.23% contamination) (Supplementary Table 1, Extended Data Fig. 9a). *E. coli* subsp. 2 and 3 are members of the B2 and B1 phylogroups respectively and have high average nucleotide identity with previously assembled[19] *E. coli* subsp. 6 and 4 (Extended Data Fig. 9b). After editing and selection, *E. coli* subsp. 2 was enriched to an average relative abundance of 99.8%, while *E. coli* subsp. 3 was enriched to 74.1% in selective liquid medium (Fig. 5a). This targeted enrichment enabled the successful assembly of genomes for the locus carrying strains, which was not possible from the pre-edit community (Extended Data Fig. 7). ET-Seq was used to map insertion locations following enrichment and showed that 100% and 99.2% of insertions occurred within 20 bp of the expected insertion site (Fig. 5b) within *E. coli* subsp. 2 and 3, respectively (Fig. 5c-d). Furthermore, we isolated both subsp. 2 and 3 by selection on solid medium to confirm on-target, clonal DART insertions by PCR and Sanger sequencing (Methods and Extended Data Fig. 10). In this way, traits of interest can be enriched within, or isolated from, a complex natural community by changing only a 32 bp guide RNA sequence in the VcDART vector.

Discussion

We have demonstrated programmable organism- and locus-specific genome editing within microbial communities, providing a new approach to microbial genetics and microbiome manipulation for research and applications. ET-Seq revealed the genetic accessibility of individual organisms within a microbial consortium. To allow for targeted editing in the community context, we created conjugative all-in-one vectors encoding two naturally occurring CRISPR-Cas transposon systems. Comparison of their on-target efficiencies showed that only one of the two systems, which we termed VcDART, enabled precise RNA-programmable microbial genome editing. VcDART accurately integrated distinct genetic payloads into the genomes of members of the synthetic soil community and the infant gut community as measured by ET-Seq. Selection of these edits allowed for enrichment and isolation of the targeted strains. Furthermore, VcDART tagging and ET-Seq tracking was used to facilitate in-community fitness measurements of genetic mutants without selection markers. Together, these tools allow for assaying a community for genetic accessibility, conducting targeted genome editing within it, and applying the resultant edits to better understand the community.

We expect community editing in the gut microbiota demonstrated here will lead to health relevant applications. The presence or absence of certain genes within a species can be the differentiator between pathogenic and commensal bacteria within the gut[23]. However, using short-read sequencing to resolve the genome of a specific trait carrying strain from a complex mixture can be confounded by genomic similarities between strains[24]. Here, using targeted editing we have shown that specific strains can be isolated out of the community and high quality genomes assembled on the basis of clinically relevant genes by programming only a 32 bp guide. The ability to shift a community towards strains with desired loci will likely also be of more applied medical importance in the future. Furthermore, the toolset to make and measure targeted edits in the gut community should facilitate fitness-tracking of edited strains, such as that demonstrated in the synthetic soil community (Fig. 3c). Genes important for virulence can now be probed for their fitness impacts within the gut microbiota. The tools developed here should facilitate genetics for understanding and clinical applications in more representative isolates and communities from the gut.

For new uncharacterized communities, draft genomes will be required to utilize the ET-Seq pipeline. However, given that genome-resolved metagenomics can now successfully reconstruct high quality draft genomes from environments containing 100s - 1000s of species[25], and because ET-Seq can map insertion sites on even highly fragmented references, this requirement should not prove limiting. In more complex communities, ET-Seq may not be able to characterize insertion efficiencies in rare or marginally genetically accisable members. For example, an organism making up 1% of its community, that had an insertion efficiency below 0.1%, may end up below the reproducibly measurable range by ET-Seq (Fig. 1b). This could limit the applicability of ET-Seq in low abundance organisms. There are, however, many enigmatic branches of the tree of life, such as candidate phyla radiation (CPR), that are abundant within communities, have no known genetic tractability, and are unisolated[26].

Despite these challenges, in more complex and uncharacterized communities, the direct measurement of insertions allowed by ET-Seq should prove especially beneficial. Previous techniques for experimentally measuring horizontal gene transfer used proxy measurements such as fluorescent markers combined with cell sorting[4–9]. These approaches provide potential for false positives and negatives because of autofluorescence of community members and variable marker expression. Furthermore, the direct sequencing of insertions allows for the retrieval of information coded in these insertions. Currently barcodes enable differentiation and tracking of unique insertion events, but in the future they could additionally mark the identity of their parent vector so a pooled library of vectors with different promoters and transposases could be tested for efficiency across diverse organisms in a single community editing experiment. A similar strategy has been used to increase efficiency of insertions in isolates[27]. In the future, we plan to apply ET-Seq alongside metagenomic sequencing to map the genetic accessibility of communities alongside their genetic content.

The information provided by ET-Seq on genetically accessible community members, the ideal delivery approach for each, and, in the future, the best delivery vector among a pool, will inform VcDART editing. Currently VcDART combined with ET-Seq can be used for testing the importance of a targeted gene in the community context (Fig. 3). Future improvements in both editing efficiency and ET-Seq limit of detection will allow further sensitivity and enable assaying the fitness of multiple genes in one experiment, as well as understanding the impact of such genes on the fitness of other organisms in the community. An advantage and disadvantage of these fitness assays, shared with pooled mutant screens[28], is that the community is mostly wild-type at the locus of interest. This allows the effect of mutations to be observed in a more natural context, but mutations that require ubiquity to impact survival, such as a mutation removing a shared molecule like a siderophore, will not be measured. VcDART can also be used to enrich a targeted strain for isolation using an antibiotic resistance marker or the creation of a metabolic niche (Fig. 4-5). While maintenance of such selective markers will be important for many applications, removal of these markers can be achieved by encoding recombinase recognition sites flanking marker cassettes to facilitate subsequent recombinase-mediated marker excision[11]. Increased editing efficiencies, improved detection sensitivity, and more universal selection methods will expand biological questions that can be answered with these tools and allow manipulation of agricultural, industrial, and health-relevant microbiomes.

Traditionally, the combined steps of culturing an environmental microbe, determining the ideal means to transform it, and implementing targeted editing could take years or could fail altogether[29]. ET-Seq together with VcDART can decrease this process to weeks and move it into the more realistic and information rich context of communities. Together, these tools decrease the need for isolation as a prerequisite for genetics and provide technologies that are essential for the new field of *in situ* genetics.

References

1. Steen, A. D. *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.* (2019).
2. Pascual-García, A., Bonhoeffer, S. & Bell, T. Metabolically cohesive microbial consortia and ecosystem functioning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, (2020).
3. Fux, C. A., Shirtliff, M., Stoodley, P. & Costerton, J. W. Can laboratory reference strains mirror 'real-world' pathogenesis? *Trends Microbiol.* **13**, 58–63 (2005).
4. Pukall, R., Tschäpe, H. & Smalla, K. Monitoring the spread of broad host and narrow host range plasmids in soil microcosms. *FEMS Microbiol. Ecol.* **20**, 53–66 (1996).
5. De Gelder, L., Vandecasteele, F. P. J., Brown, C. J., Forney, L. J. & Top, E. M. Plasmid Donor Affects Host Range of Promiscuous IncP-1β Plasmid pB10 in an Activated-Sludge Microbial Community. *Appl. Environ. Microbiol.* **71**, 5309–5317 (2005).
6. Musovic, S., Oregaard, G., Kroer, N. & Sørensen, S. J. Cultivation-Independent Examination of Horizontal Transfer and Host Range of an IncP-1 Plasmid among Gram-Positive and Gram-Negative Bacteria Indigenous to the Barley Rhizosphere. *Applied and Environmental Microbiology* vol. 72 6687–6692 (2006).
7. Musovic, S., Klümper, U., Dechesne, A., Magid, J. & Smets, B. F. Long-term manure exposure increases soil bacterial community potential for plasmid uptake. *Environ. Microbiol. Rep.* **6**, 125–130 (2014).
8. Klümper, U. *et al.* Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9**, 934–945 (2015).
9. Ronda, C., Chen, S. P., Cabral, V., Yaung, S. J. & Wang, H. H. Metagenomic engineering of the mammalian gut microbiome in situ. *Nat. Methods* **16**, 167–170 (2019).

10. Farzadfard, F., Gharaei, N., Citorik, R. J. & Lu, T. K. Efficient Retroelement-Mediated DNA Writing in Bacteria. *bioRxiv* (2020).

11. Vo, P. L. H. *et al.* CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-00745-y.

12. Hsu, B. B., Way, J. C. & Silver, P. A. Stable Neutralization of a Virulence Factor in Bacteria Using Temperate Phage in the Mammalian Gut. *mSystems* **5**, (2020).

13. Hsu, B. B. *et al.* In situ reprogramming of gut bacteria by oral delivery. *Nat. Commun.* **11**, 5030 (2020).

14. Sheth, R. U., Cabral, V., Chen, S. P. & Wang, H. H. Manipulating Bacterial Communities by in situ Microbiome Engineering. *Trends Genet.* **32**, 189–200 (2016).

15. Wu, L. R., Chen, S. X., Wu, Y., Patel, A. A. & Zhang, D. Y. Multiplexed enrichment of rare DNA variants via sequence-selective and temperature-robust amplification. *Nature Biomedical Engineering* **1**, 714–723 (2017).

16. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).

17. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

18. Petassi, M. T., Hsieh, S.-C. & Peters, J. E. Guide RNA categorization enables target site choice in Tn7-CRISPR-Cas transposons. *bioRxiv* (2020).

19. Lou, Y. C. *et al.* Infant gut strain persistence is associated with maternal origin, phylogeny, and functional potential including surface adhesion and iron acquisition. *bioRxiv* (2021).

20. Picard, B. *et al.* The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).

21. Viladomiu, M. *et al.* Adherent-invasive E. coli metabolism of propanediol in Crohn’s disease regulates phagocytes to drive intestinal inflammation. *Cell Host Microbe* **29**, 607–619.e8 (2021).

22. Dogan, B. *et al.* Inflammation-associated adherent-invasive Escherichia coli are enriched in pathways for use of propanediol and iron and M-cell translocation. *Inflamm. Bowel Dis.* **20**, 1919–1932 (2014).

23. Leimbach, A., Hacker, J. & Dobrindt, U. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* **358**, 3–32 (2013).

24. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

25. Diamond, S. *et al.* Mediterranean grassland soil C-N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nat Microbiol* **4**, 1356–1367 (2019).

26. He, C. *et al.* Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat Microbiol* **6**, 354–365 (2021).

27. Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *mSystems* **3**, (2018).

28. Cain, A. K. *et al.* A decade of advances in transposon-insertion sequencing. *Nat. Rev. Genet.* **21**, 526–540 (2020).

29. Laurenceau, R. *et al.* Toward a genetic system in the marine cyanobacterium Prochlorococcus. *Access Microbiology* **6**, 23 (2020).

30. Adler, B. A. *et al.* Systematic Discovery of Salmonella Phage-Host Interactions via High-Throughput Genome-Wide Screens. *bioRxiv* (2021).

31. Egbert, R. G. *et al.* A versatile platform strain for high-fidelity multiplex genome editing. *Nucleic Acids Res.* **47**, 3244–3256 (2019).

32. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).

33. Kalvari, I. *et al.* Non-coding RNA analysis using the rfam database. *Curr. Protoc.*

Bioinformatics **62**, e51 (2018).

34. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
35. Liu, H. *et al.* Functional genetics of human gut commensal Bacteroides thetaiotaomicron reveals metabolic requirements for growth across environments. *Cell Rep.* **34**, 108789 (2021).
36. Devon, R. S., Porteous, D. J. & Brookes, A. J. Splinkerettes--improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res.* **23**, 1644–1645 (1995).
37. Barquist, L. *et al.* The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics* **32**, 1109–1111 (2016).
38. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
39. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
40. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
41. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
42. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
43. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genom* **4**, (2018).
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

EMBnet.journal **17**, 10–12 (2011).

45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

46. Zhao, L., Liu, Z., Levy, S. F. & Wu, S. Bartender: a fast and accurate clustering algorithm to count barcode reads. *Bioinformatics* **34**, 739–747 (2018).

47. Costello, M. *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).

48. Team, R Core. R: A language and environment for statistical computing. https://www.R-project.org/ (2020).

49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

Methods

Plasmid construction and barcoding

For ET-Seq measurement of genetic tractability in community members, DNA containing a non-targeted *mariner* transposon was delivered. The *mariner* transposon integrates into "TA" sequences in recipient genomes. For delivery of the *mariner* transposon, we used the pHLL250 vector, which contains an RP4 origin of transfer (oriT), AmpR, conditional (*pir*$^+$-dependent) R6K origin, and an AseI restriction site to facilitate depletion of vector from DNA samples in ET-Seq library preparations[30]. Unique to each transposon on this vector is a random 20 bp barcode sequence to aid in the discrimination of unique insertion events from duplications of the same insertion due to cell division or PCR. The pHLL250 vector contains greater than 10 million barcode variants.

DART vectors were designed to encode all components required for delivery and editing (Supplementary Table 3 and Extended Data Fig. 4). VcCasTn genes, crRNA, and transposon were synthesized as gBlocks (IDT). pHelper_ShCAST_sgRNA was a gift from Feng Zhang (Addgene plasmid #127921; http://n2t.net/addgene:127921; RRID:Addgene_127921) and was used to clone ShCasTn genes and sgRNA. pDonor_ShCAST_kanR was a gift from Feng Zhang (Addgene plasmid # 127924 ; http://n2t.net/addgene:127924 ; RRID:Addgene_127924) and was used to clone the ShCasTn transposon. *tns* genes, *cas* genes, and crRNA/sgRNA were consolidated into a single operon (with various promoters and transcriptional configurations) on the same vector as the cognate transposon. The left end of the cognate transposon was encoded downstream of the crRNA/sgRNA, followed by cargo, barcode, and transposon right end. DART transposon LE and RE were designed to include the minimal sequence that both included all putative TnsB binding sites and was previously shown to be functional[16,17]. Specifically, VcDART LE (108 bp) and RE (71 bp) each encompass three 20 bp putative TnsB binding sites, spanning from the edge of the 8 bp terminal ends to the edge of the third putative TnsB binding site[17].

ShDART LE (113 bp) spans the boundaries of the long terminal repeat and both additional putative TnsB binding sites, while the RE (211 bp) encompasses the long terminal repeat and all four additional putative TnsB binding sites[16].

Vectors were cloned using BbsI (NEB) Golden Gate assembly of part plasmids, each encoding different regions of the final plasmid. The constitutive *lacZY* cassette, amplified from *E. coli* MG1655 genomic DNA with strong constitutive promoter BBa_J23119 appended to the 5' end, was inserted into the cargo region of BbsI-assembled VcDART vectors in a subsequent step using LguI (NEB) Golden Gate assembly. Of note, the backbone encodes RP4 oriT, AmpR, conditional R6K origin, and an AsiSI+SbfI double digestion site for vector depletion during ET-Seq library preparations. A 2xBsaI spacer placeholder enabled spacer cloning with BsaI (NEB) Golden Gate. A 2xBsmBI barcode placeholder was encoded immediately inside the transposon right end and was used for barcoding as described below. Part plasmids were propagated in *E. coli* Mach1-T1R (QB3 Macro Lab). Golden Gate reactions for all-in-one vector assembly were purified with DNA Clean & Concentrator-5 (Zymo Research) and electroporated into *E. coli* EC100D-*pir*+ (Lucigen).

DART vectors to be assayed by ET-Seq were barcoded by BsmBI (NEB) Golden Gate insertion of random barcode PCR product into the 2xBsmBI barcode placeholder using a previously reported method[27] with slight modifications. A 56-nt ssDNA oligonucleotide encoding a central tract of 20 degenerate nucleotides (oBFC1397) was amplified with BsmBI-encoding primers oBFC1398 and oBFC1399 using Q5 High-Fidelity 2X Master Mix (NEB) in a six-cycle PCR (98°C for 1 min; six cycles of 98°C for 10 s, 58°C for 30 s, and 72°C for 60 s; and 72°C for 5 min). Barcoding Golden Gate reactions were purified with DNA Clean & Concentrator-5. To remove residual non-barcoded vector, reactions were digested with 15 U BsmBI at 55°C for 4 hr, heat inactivated at 80°C for 20 min, treated with 10 U Plasmid-Safe ATP-Dependent DNase (Lucigen) exonuclease at 37°C for 1 hr, heat inactivated at 70°C for 30 min, and purified with DNA Clean & Concentrator-5.

Randomly barcoded conjugative vectors were electroporated into *E. coli* EC100D-*pir*+, followed 1 hr recovery in 1 mL pre-warmed SOC (NEB) at 37°C 250 rpm, serial dilution and spot plating on LB agar plus 100 μg mL^{-1} carbenicillin to estimate library diversity, and plating the full transformation across 5 LB agar plates containing carbenicillin (and other appropriate antibiotics when transposon cargo contained other resistance cassettes). To prepare barcoded conjugative vector plasmid stock, all 5 agar plates were scraped into a single pool and midiprepped (Zymo Research). All conjugations were performed using the diaminopimelic acid (DAP) auxotrophic RP4 conjugal donor *E. coli* strain WM3064. Donor strains were prepared by electroporation with 200 ng barcoded vectors, followed by recovery in SOC plus DAP at 37°C and 250 rpm and inoculation of the entire recovery culture into 15 mL LB containing DAP and carbenicillin in 50 mL conical tubes, followed by overnight cultivation at 37°C and 250 rpm. Donor serial dilutions were spot plated on LB agar plus carbenicillin to estimate final barcode diversity.

Guide RNA design

In all experiments, VcCasTn gRNAs used 32 nt spacers and a 5'-CC Type IF PAM, while ShCasTn gRNAs used 23 nt spacers and a 5'-GTT Cas12k PAM. All gRNAs were designed to bind in the first half of the target CDS to ensure functional knockout by transposon insertion (Supplementary Table 4). Safe Site gRNAs were designed to bind at least 100 bp inside Safe Site boundaries as specified below. Fimbriae and propanediol utilization locus-targeting gRNAs were designed to target intergenic sites near one end of the predicted unique gene clusters. After in-community editing, selective enrichment, sequencing, and *de novo* assembly of the *E. coli* subsp. 3 genome, we determined that the enriched target site contained a 1 bp mismatch relative to the designed gRNA that was not observed in the reference genome. Successful enrichment of *E. coli* subsp. 3 indicated that this single mismatch did not prevent transposition. Off-target potential for all gRNAs was assessed using BLASTn (-dust no -word_size 4) of spacers against a local BLAST database created from all genomes present in an experiment, and spacers were discarded if off-target hits with E-value < 15 were identified. gRNAs with less seed region complementarity to off-targets were prioritized. Non-targeting gRNAs were designed by scrambling the spacer until no significant matches were found.

Identification of Safe Sites for targeted genome integration

Putative safe genome integration loci, referred to as Safe Sites, were identified in *K. michiganensis* (GCF_002090205.1) and *P. simiae* (GCF_000698275.1) following previously reported methods[31]. Specifically, all intergenic regions between two convergently transcribed genes were rank ordered by size and filtered to remove those containing predicted RNA features[32,33], those inside or adjacent to a putative mobile genetic element, and those flanked by at least one likely essential gene (in which no insertions were obtained in a high coverage genome-wide transposon mutant library) or genes exhibiting fitness defects in any previously screened conditions[34]. The longest intergenic region fulfilling these criteria was selected as the Safe Site for VcDART integration, specifically coordinates 3,533,769-3,534,285 in *K. michiganensis* and 3,209,633-3,210,436 in *P. simiae*.

Delivery methods

For natural transformation and electroporation, a culture of the community or isolate to be transformed was subcultured at OD_{600} = 0.2 and grown to OD_{600} = 0.5. For natural transformation 200 ng of vector harboring the *mariner* transposon (pHLL250[30]) for non-targeted insertion, or water for the negative control were added to 4 mL of OD_{600} = 0.5 outgrowth. Cultures were incubated for 3 hours shaking at 250 rpm at temperature appropriate for the isolate or community before being moved to the appropriate downstream analysis.

For electroporation, 20 mL of the community or isolate at OD_{600} = 0.5 was put on ice, centrifuged at 4,000*g* at 4°C for 10 minutes, and washed four times with 10 mL sterile ice-cold Milli-Q H_2O. After a final centrifugation the pellet was resuspended in 100 µL of 2 ng/µL vector (pHLL250 or VcDART), or 100 µL of water as a negative control. This solution was then pipetted into a 0.2 cm gap ice-cold cuvette and electroporated at 3 kV, 200Ω, and 25 µF. The cells were immediately recovered into 10 mL of the community's or isolate's preferred medium and incubated shaking for 3 hours before being moved to the appropriate downstream analysis.

For conjugation, *E. coli* strain WM3064 containing the *mariner* transposon (pHLL250) for non-targeted editing, or the VcDART for targeted editing, was cultured overnight in LB supplemented with carbenicillin (100 µg/mL) and DAP (60 µg/mL) at 37°C. Before conjugation the donor strain was washed twice in LB (centrifugation at 4,000*g* for 10 minutes) to remove antibiotics. Then, the equivalent of 1 mL of culture at OD_{600} value 1 (1 OD_{600}*mL) of the donor was added to 1 OD_{600}*mL of the recipient community or isolate and the mixture was plated on a 0.45 µm mixed cellulose ester membrane (Millipore) topping an agar plate of the recipient's preferred media without DAP. Plates were incubated at the preferred temperature for the recipient community or isolate for 12 hours before the growth was scraped off the filter into the media of the recipient community or isolate for downstream analysis.

ET-Seq library preparation

The insertion junction sequencing library prep strategy for ET-Seq can be used (modification may be necessary) in any circumstance where high efficiency mapping of inserted DNA to a host loci is desired. For our purposes, DNA of the edited community or isolate was first extracted using the DNeasy PowerSoil Kit (QIAGEN). Five hundred ng of DNA was used each for insertion junction sequencing and metagenomic library prep. As an internal standard, DNA from a previously constructed mutant library of *Bacteroides thetaiotaomicron* VPI-5482[35], a species not present in the synthetic soil member community, was spiked into the community DNA at a ratio of 1/500 by mass. The *B. thetaiotaomicron* library had undergone antibiotic selection for its transposon insertions and was thus assumed to represent 100% transformation efficiency (i.e. every genome contained at least one *mariner* transposon insertion).

For metagenomic sequencing, library prep was conducted by the standard ≥100 ng protocol from the NEBNext Ultra II FS DNA Library Prep Kit for Illumina (NEB). For insertion junction sequencing, the same protocol was used with a number of modifications enumerated here (Extended Data Fig. 1). This insertion junction sequencing protocol has also been tested successfully with the ≤ 100 ng protocol of the NEBNext Ultra II FS DNA Library Prep Kit (NEB) and the KAPA HyperPlus Kit (Roche). For fragmentation an 8 minute incubation was used. A custom splinklerette adaptor was used during adaptor ligation to decrease non-specific amplification (Supplementary Table 5)[36,37]. For size selection 0.15X (by volume) SPRIselect (Beckman Coulter, Cat # B23318) or NEBNext Sample Purification Beads (NEB) were used for the first bead selection and 0.15X (by volume) were added for the second. From this selection, the DNA was eluted in 44 µL (instead of the suggested 15 µL) where it undergoes digestion before enrichment to cleave intact transposon delivery vector. All bead elutions were performed with Sigma Nuclease-Free water. pHLL250 underwent AseI digestion, while DART vectors underwent double digestion by AsiSI and SbfI-HF (NEB) (Supplementary Table 3). The DNA then underwent a sample purification using 1X AMPure XP beads (Beckman Coulter) to prepare it for PCR enrichment.

In PCR enrichment, the transposon junction was amplified by nested PCR. The PCRs followed the NEBNext Ultra II FS DNA Library Prep Kit for Illumina (NEB) PCR protocol, however in the first PCR the primers were custom to the transposon and the adaptor and the PCR was run for 25 cycles (Supplementary Table 5). The enrichment then underwent sample purification with a 0.7X size selection using SPRIselect or NEBNextSample Purification Beads from which 15 µL were eluted for the second PCR. This second PCR used custom unique dual indexing primers specific to nested regions of the insertion and adaptor and 6 cycles were used (Supplementary Table 5). Then another 0.7x size selection was conducted and the final library was eluted in 30 µL. Samples for metagenomic sequencing and insertion junction sequencing were then quality controlled and multiplexed using 1X HS dsDNA Qubit (Thermo Fisher) for total sample quantification, Bioanalyzer DNA 12000 chip (Agilent) for sizing, and qPCR (KAPA) for quantification of sequenceable fragments. Samples were sequenced on the iSeq100, HiSeq4000, and NovaSeq 6000 platforms.

Genome sequencing, assembly, taxonomic classification, and database construction

For a full list of genome sequences used as read mapping references in this study see Supplementary Table 1. For synthetic soil community genomes assembled as part of this study, cultures were grown on R2A medium for 24 hours at 30°C and genomic DNA was extracted with the DNeasy Blood and Tissue DNA Kit (Qiagen) with pre-treatment for Gram-positive bacteria. Genomic DNA was sheared mechanically with the Covaris S220 and processed with the NEBNext DNA Library Prep Master Mix Set for Illumina (NEB) before submitting for sequencing on an Illumina MiSeq platform generating paired end 150 bp reads. Raw sequencing reads were processed to remove Illumina adapter and phiX sequence using BBduk with default parameters, and quality trimmed at 3' ends with Sickle using default parameters (https://github.com/najoshi/sickle). Assemblies were conducted using IDBA-UD v1.1.1[38] with the following parameters: –pre_correction –mink 30 –maxk 140 –step 10. Following assembly, contigs smaller than 1 kbp were removed and open reading frames (ORFs) were then predicted on all contigs using Prodigal v2.6.3[39]. 16S ribosomal rRNA genes were predicted using the 16SfromHMM.py script from the ctbBio python package using default parameters (https://github.com/christophertbrown/bioscripts). Transfer RNAs were predicted using tRNAscan-SE[40]. The full metagenome samples and their annotations were then uploaded into our in-house analysis platform, ggKbase, where genomes were manually curated via the removal of contaminating contigs based on aberrant phylogenetic signatures (https://ggkbase.berkeley.edu). For the infant gut community, the reference genomes used for mapping and analysis were constructed and described previously[19]. Genome recovery from gut enrichment samples, assemblies of shotgun metagenomic data were conducted as above and automatic genome binning was performed as previously described[19]. Manual curation of insertion loci was performed for assemblies of *E. coli* subsp. *2* and subsp. *3* as described in [41]. The dRep dereplicate pipeline (v3.2.2)[42] was run using default program parameters to both dereplicate *E. coli* genomes using Average Nucleotide Identity (ANI) and assess the quality of genomes directly recovered from the gut enrichment samples. *E. coli* phylotyping was performed using the ClermonTyping tool using default program parameters[43].

For each ET-Seq experiment a genomic database is constructed using the ETdb component of the ETsuite software package. Each database contains the nucleotide sequences of the expected organisms in a sample, any vectors used, any conjugal donor, and the spike-in control organism. Briefly, all genomic sequences are formatted into a bowtie2 index to allow read mapping, a tabular correspondence table between all scaffold names and their associated genome is constructed (scaff2bin.txt), and a table (genome_info.txt) of standard genomic statistics is calculated including genome size, GC content, and number of scaffolds. Following database construction, a label is manually added to each entry in the genome info table to indicate if the entry represents a target organism, a vector, or a spike-in control organism. All data are propagated into a single folder that can be used by the ETmapper software for downstream mapping and analysis.

Identification and quantification of insertion junctions and barcodes

To identify and map transposon insertion junctions and their associated barcodes in a mixed population of microbial cells, reads (150 bp X 2) generated from PCR amplicons of putative transposon insertion junctions are first processed using the ETmapper component of the ETsuite software package implemented in R with the following steps: First reads are quality trimmed at the 3' end to remove low quality bases (Phred score ≤ 20) and sequencing adapters using Cutadapt v2.10[44]. Cutadapt is then used to identify and remove provided transposon model sequences from the 5' end of forward reads, requiring a match to 95% of the shortest transposon sequence in a provided set and allowing a 2% error rate. Read pairs where no transposon model sequence is identified in the forward read are discarded. All identified and trimmed transposon models are paired with their respective reads, stored, and barcodes are identified in these sequences by searching for a known primer binding site sequence flanking the 5' end of the barcode (5'-CTATAGGGGATAGATGTCCACGAGGTCTCT-3') allowing for 1 mismatch. Subsequently, the 20 bp region following the known primer binding site is extracted as the barcode sequence and associated with its respective read. The 3' end of the paired reverse reads are then trimmed to remove any transposon model sequence using Cutadapt, and only read pairs where one mate is at least ≥ 40 bp following all trimming are retained for downstream mapping and analysis. The fully trimmed paired end reads now consisting of only genomic sequence following the transposon insertion site are mapped to the ETdb database used in a given experiment using bowtie2[45]. with the following options: --rdg 60,3"," --rfg 60,3 to disallow any gaps in the alignment Mapped read files are converted into a hit table indicating the mapped genome, scaffold, genomic coordinates, mapQ score, and number of alignment mismatches, number of alignment mismatches in the first 3bp of the alignment, and last 5bp that were present in the identified transposon model sequence for each read pair using a custom Python script, bam_pe_stats.py, provided with ETsuite. This table is then merged with read-barcode assignments to generate a final hit table with the mapping information about each read pair, the transposon model identified, and the associated barcode found for that read pair. Mapped read pairs are only retained for downstream quantification if both reads map to the same genome, at least one mapped read in a pair has a mapQ score ≥ 20, and a barcode was successfully identified and associated with the read pair. Finally putative chimeric sequences are filtered by examining the transposon model - genome junction within forward reads. Read pairs are removed from the analysis if in the forward read the last 5bp of the transposon model is not an exact match to the predicted mode for that read pair or if there are any alignment mismatches between first 3bp of the genomic sequence in the forward read and the genome sequence that read was mapped to. To quantify the number of unique barcodes and their associated reads mapping to organisms in each sample of an experimental run, the filtered hit tables were processed using the ETstats component of the ET-Seq software package with the following steps: Initially, all barcodes identified across all samples in an experiment are aggregated and clustered using Bartender[46] with the following supplied options: -l 4 -s 1 -d 2. Barcode clusters and their associated barcodes/reads were only retained if all of the following criteria were true: (1) ≥ 75% of the reads in a cluster mapped to one genome (the majority genome), (2) ≥ 75% of the reads in a cluster were associated with the same transposon model (the majority model), and (3) the barcode cluster had at least 2 reads. Subsequently, when quantifying reads and barcodes in each sample of an experiment, the genome a read was mapped to and the transposon model it was associated with had to agree with the majority assignments for the barcode cluster assigned to that read's barcode to be counted. Finally, we were aware that Illumina patterned flow cell related index swapping would result in reads from a barcode cluster being misassigned across samples, even when using unique dual indexing[47]. We could not simply limit barcode clusters to be associated with only one sample, as our spike in control organisms contain the same pool of barcodes and are added to every sample. Thus we estimated an empirical index swap rate across each experiment and required that the number of reads (X) for a barcode to be positively identified in a sample be always ≥ 2 and ≥ the binomial mean of observed read counts expected in any sample for a barcode cluster with (R) reads across (N) samples based on the estimated swap rate (S) + 2 standard deviations (**Eqn. 1**)

Eqn. 1: $X \geq \left(R \times \left(\frac{S}{N}\right)\right) + 2 \times \sqrt{R \times (1 - S) \times S} \;\; \& \;\; X \geq 2$

The index swap rate for an experiment was empirically estimated from barcode clusters assigned only to target organisms based on the assumption that it would be highly unlikely for a barcode cluster to have truly originated from independent integration events into the same organism in more than one sample. Thus we assumed that for each barcode cluster associated with target organisms, the majority of reads originated from the true sample and reads assigned to other samples represented swaps. This is opposed to barcode clusters associated with our spike-in organism, conjugal donor organism, or vectors which contain the same pool of barcodes directly added to multiple samples. To identify swapped read counts we first quantify the total count of all reads assigned to the majority genome across barcode clusters but that are not associated with the majority sample of that cluster (E). Then we quantify the total count of reads associated with the majority genome and associated with the majority sample across all clusters (C). Then experiment wide swap rate was estimated by dividing the total number of reads not associated with majority samples by the total number of reads (**Eqn. 2**)

Eqn. 2: $S = \frac{E}{(E + C)}$

Following filtering, a hit table is returned that indicates for each genome in each sample, the number of unique barcode clusters that were recovered, and the total number of reads associated with these barcodes.

Metagenomic data processing and coverage calculation

Each ET-Seq sample is split and in parallel undergoes shotgun metagenomic sequencing to determine the relative quantities of organisms present in the sample at the time of sampling. Raw read files from metagenomic data are also processed using the ETmapper component of the ETsuite software package with the following steps: First reads are quality trimmed at the 3' end to remove low quality bases (Phred score ≤ 20) and sequencing adapters using Cutadapt v2.10[44]. Read pairs where at least one mate is not ≥ 40 bp in length are discarded. Trimmed read pairs are mapped to the ETdb database used in a given experiment using bowtie2[45] with default parameters. Mappings are filtered to require a minimum identity ≥ 95% and minimum mapQ score ≥ 20, and coverage is calculated using a custom script, calc_cov.py, included with the ETsuite software.

ET-Seq normalization and calculation of insertion efficiency

To account for differences in sequencing depth, transposon junction PCR template amount, and relative abundance of microbes in a community the data generated from both ET-Seq and shotgun metagenomics were each normalized independently to values from the spike in control organism, *B. thetaiotaomicron,* and then ET-Seq data is subsequently normalized by metagenomic abundance as follows: Initially read count tables from ET-Seq and metagenomics are filtered to remove any ET-Seq read count associated with < 2 barcodes and any metagenomic read count < 10 reads. Next a size factor for each sample is calculated based on the geometric mean of *B. thetaiotaomicron* reads for ET-Seq samples and *B. thetaiotaomicron* coverage for metagenomics samples. ET-Seq read counts and metagenomic coverage values are then divided by their respective sample size factors to create normalized values. Normalized ET-Seq read counts are then divided by their paired normalized metagenomic coverage values to generate ET-Seq read counts that are fully normalized to both ET-Seq sequencing depth and metagenomic coverage. Finally fully normalized ET-Seq read counts for target organisms are divided by the fully normalized ET-Seq read count of *B. thetaiotaomicron* from an experiment (a constant that represents the number of reads that would be obtained from an organism with 100% of its chromosomes carrying insertions). The resulting values for each target organism in a sample represent an estimate of the fraction of that organism's population that received insertions (Insertion Efficiency). Additionally, we multiply a target organism's insertion efficiency by the fractional relative abundance of that organism in a sample, based on metagenomic data, to estimate the fraction of an entire sample population that is made up of cells of a given species that received insertions (Insertion-Receiving Fraction in Total Community).

ET-Seq validation

To validate ET-Seq and gain understanding of both the relationship of our assay outputs to known populations of edited cells and the limits of the assay, a library of *K. michiganensis* transposon mutants was constructed by antibiotic selection following conjugation with pHLL250 (as described above), and this library was added to untransformed samples of the synthetic soil community to create a transformed cell concentration gradient. Triplicate samples were created where 1%, 0.1%, 0.01%, 0.001% and 0% of the total *K. michiganensis* cells (by OD_{600}) in the mixture were those derived from the transformed library. All samples (n = 15) were subjected to ET-Seq (as described above), and pooled samples across all concentrations for each technical triplicate (n = 3; 5 concentrations) were analyzed for community composition using shotgun metagenomics (as described above). Additionally, to derive the fraction of transformed *K. michiganensis* cells that made up the total community (not just the *K. michiganensis* sub-population), the known fraction of *K. michiganensis* cells that were transformed in a sample was multiplied by the measured relative abundance of *K. michiganensis* in a given technical replicate, and these values were averaged across technical replicates.

A log-log linear regression was performed using the lm function in the base package of R[48] using the known fraction of transformed *K. michiganensis* cells that made up the total community as the independent variable and the ET-Seq estimated per community insertion efficiency as the dependent variable. The sample where transformed *K. michiganensis* made up 0% of the community was not included nor was a single experimental sample where ET-Seq recovered an edited fraction of zero.

Chimera measurements

To test whether chimeras between delivery vector and wildtype DNA occur during library preparation, various quantities of delivery vector were spiked into unmodified DNA directly before library preparation. We pooled 291 ng DNA from the wild type synthetic soil community with 209 ng wild type *Sinorhizobium meliloti 1021* DNA, 1 ng insertion containing *B. thetaiotaomicron* internal standard, and 0.001-10ng donor vector (pHLL250) depending on sample. The quantity of *S. meliloti* DNA was chosen to be in similar relative proportion to the abundant community member *K. michiganensis.* Vector spike-in quantities were chosen to be centered around the estimated amount of DNA coming from the vector in a real ET-Seq experiment. This mixture underwent standard ET-Seq library prep and sequencing (described in ET-Seq library preparation section). Insertions to *S. meliloti* were used as a signal for chimeric reads (Extended Data Fig. 2).

Multiple delivery experiments in communities

To test multiple delivery methods on the synthetic soil community, all members were grown at 30°C with *Bacillus sp.* AnTP16 and *Methylobacterium sp.* AMD150 in R2A liquid media while all other members were inoculated in LB. Equal amounts of community members were then combined by OD_{600}. This consortium then underwent transformation (of pHLL250), conjugation (pHLL250 in WM3064), and electroporation of the pHLL250 vector (described in Delivery methods section). After delivery the community was spun down at 5,000*g* for 10 minutes, washed once with LB and then spun down and frozen at -80°C until genomic DNA extraction.

Benchmarking DART systems in *E. coli*

We first constructed several DART systems to identify variants capable of efficient transposition by conjugative delivery to *E. coli*. We performed parallel conjugation of each DART vector variant containing Gm^R transposon cargo (2.1 kbp) and either a non-targeting gRNA or one of two *lacZ*-targeting gRNAs for each system. For VcDART, variation of the promoter controlling the expression of VcCasTn components did not significantly impact transposition efficiency (Extended Data Fig. 4c-d). Similarly for ShDART, expression of the sgRNA in three distinct transcriptional configurations did not significantly impact transposition efficiency (Extended Data Fig. 4e-f). These distinct ShDART sgRNA transcriptional configurations were tested to determine if 5' and 3' sgRNA ends were critical for function, yet all configurations achieved similar editing efficiencies with both *lacZ*-targeting and non-targeting sgRNAs, in line with the previously documented off-target ShCAST Tn7 activity observed in absence of Cas12k[16]. For VcDART, Cas6 catalyzes processing of its crRNA, so we deemed similar optimization of the 5' and 3' ends of its crRNA as unnecessary. Since promoter and transcriptional configuration variation had insignificant effects on transposition efficiency – and to remove the requirement for promoter induction and reliance on T7 RNA polymerase as well as to present the best on-/off-target comparison to Strecker et al. – we performed target specificity benchmarking of VcDART and ShDART using the same constitutive P_{lac} promoter derived from pHelper_ShCAST_sgRNA[16]. In this experiment, ShDART Cas and transposase genes and sgRNA were encoded in the original transcriptional configuration and under control of the same promoter in which ShCasTn was first characterized in pHelper_ShCAST_sgRNA by Strecker et al.[16].

The *lacZ*-targeting gRNAs were designed to target the *lacZ* α-peptide present in the conjugation recipient strain *E. coli* BL21(DE3) but absent in the *lacZ*ΔM15 strains used as cloning host (*E. coli* EC100D-*pir*+) or conjugation donor (*E. coli* WM3064), preventing transposition until delivery into the recipient cell (Extended Data Fig. 4a). Donor WM3064 strains were transformed and cultivated as described above, and recipient BL21(DE3) was inoculated from glycerol stock into 100 mL LB in a 250 mL baffled shake flask at 37°C 250 rpm. Conjugations were performed as described above using LB medium and 37°C incubation for every step, except that 0.1 mM IPTG was added to VcDART conjugation plates in Extended Data Fig. 4d to induce transcription from $P_{T7\text{-}lac}$ and T7 RNA polymerase expression in *E. coli* BL21(DE3). Transposition efficiencies were calculated as the percentage of colonies resistant to 10 μg mL^{-1} gentamycin relative to viable colonies in absence of gentamycin.

On/off-target analysis was performed for one *lacZ*-targeting guide for each DART system by outgrowth under selection followed by genomic DNA extraction and ET-Seq. Specifically, approximately 10,000 transconjugant cfu were plated on LB agar with gentamycin, incubated at 37°C overnight, scraped from agar into liquid LB medium, diluted to OD_{600} = 0.25 into 10 mL LB plus gentamycin in 50 mL conical tubes, incubated at 37°C 250 rpm until OD_{600} = 1.0, centrifuged at 4,000*g*, and frozen for downstream analysis. To determine the percent of selectable transposed colonies possessing on-target and off-target edits, the total number of selectable colonies was adjusted (Extended Data Fig. 4b) for on-target and off-target percent as determined by ET-Seq (Fig. 2b). ET-Seq analysis was conducted on triplicate platings of DART transconjugants (n = 3 for each system) to identify transposon insertion locations and quantify on-target vs. off-target insertions. As the targeted genomic region encoding the *lacZ* α-peptide is duplicated in *E. coli* BL21(DE3), one of the two duplicated regions (749,903 bp --> 750,380 bp) was removed prior to analysis to allow unambiguous mapping assignment. Subsequently, the standard ETsuite analysis pipeline (as described above) was used to identify and map 300 bp X 2 reads containing transposon junctions back to the recipient BL21(DE3) genome and cluster barcodes that corresponded to unique insertion events. To confirm an insert location we first identified the exact transposon-genome junction mapping coordinate that was the most frequent in the reads of a barcode cluster (prime location) then required that a barcode cluster had: (1) at least 75% of its reads coming from within 3 bp of the prime location and (2) at least 75% of its reads mapping to the same strand. If these criteria were true the barcode cluster was counted as a unique insertion and the prime location was used as the mapping locus by ET-Seq. An on-target insertion was evaluated as a barcode cluster with an insertion location within 200 bp downstream of the 3' end of the protospacer target. Finally all distances reported from the protospacer target site were calculated from the last base pair of the 3' end of the protospacer.

Targeted genetic mutant fitness assay in a microbial community

Barcoded VcDART vectors encoding constitutively expressed VcCasTn, a minimal transposon cargo containing only a barcode feature for ET-Seq analysis (0.3 kbp), and a *K. michiganensis* Safe Site-targeting (barcoded pBFC0882) or *pyrF*-targeting (barcoded pBFC0883) constitutive crRNA were electroporated as described above into *E. coli* WM3064. After overnight cultivation at 37°C in LB supplemented with 100 µg mL^{-1} carbenicillin and DAP, both donors were pooled at equivalent OD$_{600}$ to generate a single donor pool containing two crRNAs. Conjugation of this donor pool into the synthetic soil community was performed as described above on filter-topped LB agar plates with 12 hr incubation at 30°C. Lawns were scraped from filters into 10 mL LB medium, vortexed, and 1 OD$_{600}$*mL from each conjugation plate resuspension was plated on LB agar supplemented with 1 mg mL^{-1} 5-FOA. The remainder of each conjugation plate resuspension was centrifuged at 4000*g* to collect cells for storage at -80°C for downstream ET-Seq analysis and shotgun metagenomic sequencing (Novogene). Following 4.5 days of incubation of LB agar plates with 5-FOA at 30°C, all cells were scraped from the agar into 15 mL LB medium and centrifuged at 4000*g* to collect cells for storage at -80°C for the same sequencing analyses.

The outputs generated by the ETstats script from the ETsuite pipeline, were additionally filtered for barcode clusters that had greater than 80% of their reads mapping to within 3bp of the most frequent mapping location for that cluster. The filtered ETstats output were then converted to a bed file format and the number of unique barcodes or reads that map to the genome within a 200bp window of the VcDART target site were identified using Bedtools [49]. For the genome-wide targeting plots (insets of Fig. 3), the respective genomes were divided into 500 bp bins and the frequency of unique barcodes from the ETstats output mapping to each bin were calculated using bedtools.

Targeted isolation of edited species in a synthetic soil community through antibiotic selection

VcDART vectors encoding constitutively expressed VcCasTn, constitutive *bla*:*aadA* transposon cargo (2.7 kbp), and either a non-targeting (pBFC0888), *K. michiganensis pyrF*-targeting (pBFC0825), or *P. simiae* WCS417 *pyrF*-targeting (pBFC0837) constitutive crRNA were transformed into *E. coli* WM3064. Conjugations of these vectors into the synthetic soil community were performed as described above on filter-topped LB agar plates with 12 hr incubation at 30°C. Lawns were scraped from filters into 10 mL LB medium, vortexed, and 1 OD_{600}*mL from each cell resuspension was plated on LB agar supplemented with 1 mg mL^{-1} 5-FOA, 100 µg mL^{-1} carbenicillin, 100 µg mL^{-1} streptomycin, and 100 µg mL^{-1} spectinomycin. Following 3.5 days of incubation at 30°C, all cells were scraped from the agar into 15 mL R2A medium, vortexed, diluted into 10 mL R2A supplemented with 20 mg mL^{-1} uracil, carbenicillin, streptomycin, spectinomycin, and either 1 mg mL^{-1} 5-FOA (*K. michiganensis*-targeting and non-targeting) or 0.8 mg mL^{-1} 5-FOA (*P. simiae*-targeting and non-targeting) to OD_{600} = 0.02, and split evenly across 4 wells (2.5 mL/well) of a 24 deep well plate. Samples were cultivated at 30°C and 750 rpm and harvested when dense growth was observed in samples treated with genome-targeting crRNAs (2 days for *P. simiae* in 0.8 mg mL^{-1} 5-FOA and 5 days for *K. michiganensis* at 1 mg mL^{-1} 5-FOA). At the time of sampling, non-targeting control cultures exhibited no growth. A small portion of these cultures was serially diluted in R2A and plated on LB agar plus antibiotics to isolate and assay colonies by targeted PCR and Sanger sequencing of *pyrF* loci. The remainder of each culture was centrifuged at 4,000*g* for 10 min and frozen at -80°C for downstream shotgun metagenomic sequencing along with pre-conjugation synthetic soil community samples (Novogene). Fractional abundance was calculated using the ET-Seq light metagenomics pipeline as described above for pre-conjugation synthetic soil community cultures and post-selection *pyrF*-targeted cultures.

Targeted enrichment of edited species in a synthetic soil community through lactose consumption

VcDART vectors encoding constitutive VcCasTn, transposon cargo containing *E. coli lacZY* under constitutive control of strong minimal promoter BBa_J23119 (4.7 kbp), and either a non-targeting (pBFC0982) or *P. simiae* Safe Site-targeting (pBFC0973) constitutive crRNA were transformed into *E. coli* WM3064. Conjugations of these vectors into the four-member synthetic soil community were performed as described above on filter-topped LB agar plates with 12 hr incubation at 30°C. Specifically, this community consisted of four members of the nine-member synthetic soil community that are unable to metabolize lactose. These four organisms were mixed in equal amounts by OD_{600} as described previously for the synthetic soil community. To harvest these conjugations, lawns were scraped from filters into 50 mL phosphate buffered saline (PBS), vortexed, centrifuged at 3000*g* for 30 min, resuspended in 15 mL PBS, and 1 OD_{600}*mL from each cell resuspension was plated on RCH2_defined_noCarbon (RCH2)[34] agar either with no carbon source added or with 25 mM beta-lactose supplementation. Following 4.5 days of incubation at 30°C, only communities treated with the *P. simiae* Safe Site-targeting guide exhibited colonies on RCH2 with 25 mM beta-lactose, while no growth was observed for any conjugations transferred onto RCH2 without a carbon source. The edge of 32 colonies from each biological replicate was picked into LB medium and cultivated overnight, followed by genomic DNA extraction, PCR amplification of the *P. simiae* genomic Safe Site junction with the VcDART *lacZY* transposon, and Sanger sequencing, confirming that 100% of picked colonies (96 in total across three biological replicates) contained on target integration of the *lacZY* cassette. Additionally, all colonies were scraped as a pool from the RCH2 agar plate containing beta-lactose into 10 mL PBS, vortexed, diluted into 3 mL RCH2 with 25 mM beta-lactose to OD_{600} = 0.02 in 14 mL round bottom culture tubes, and incubated at 30°C and 250 rpm. Liquid cultures exhibited dense growth and were harvested after 3 days by centrifugation at 4,000*g* for 10 min and frozen at -80°C for downstream shotgun metagenomic sequencing (Novogene) along with pre-conjugation four-member synthetic soil community samples.

Development of an infant gut microbiota

All handling of the infant stool microbiota was performed in an anaerobic chamber. Reagents were sparged with N_2 to remove oxygen and supplemented with 0.5 g/L L-cysteine hydrochloride as a reductant (MilliporeSigma) and 1 mL of 0.1% resazurin sodium salt (Sigma-Aldrich) added as an oxygen indicator.

Infant stool sample for the inoculum was from individual 31 on their 90th day of life, as reported in Lou et al., 2021[19]. Resuspension was conducted by inoculating 300 mg of stool into 600 µL of PBS, homogenizing by pipetting, and adding 15 µL of this mixture into 3 mL of Brain Heart Infusion (BHI) liquid medium (BD) in a 24 deep well block. The culture was allowed to recover for two days at 37°C without shaking, after which it underwent three more passages of 30 µL into 3 mL of fresh BHI liquid medium, with each allowed to grow 24 hr at 37°C. After the final passage, a 500 µL aliquot of the culture was taken for a 25% glycerol stock and the remaining 2.5 mL was harvested by centrifugation at 4,000*g* for 10 min and pellets were frozen at -80°C for subsequent DNA extraction and metagenomic sequencing of the enrichment. The glycerol stock was used as inoculum for targeted editing experiments.

Targeted enrichment of edited strains in an infant gut community through antibiotic selection

VcDART vectors encoding constitutive VcCasTn, constitutive *bla*:*aadA* transposon cargo (2.7 kbp), and either an *E. coli* subsp. 2 fimbriae locus-targeting (pBFC1050) or an *E. coli* subsp. 3 propanediol utilization locus-targeting (pBFC1046) constitutive crRNA were transformed into *E. coli* WM3064. Conjugations of these vectors into the infant gut community were performed similarly to those described above with modifications, notably that all steps were performed anaerobically. The gut enrichment community was inoculated from frozen glycerol stock into 100 mL BHI in a 250 mL flask and incubated in a stationary incubator at 37°C for 36 hr. Conjugal donor strains grown aerobically, were washed twice with BHI, and resuspended in anaerobic BHI supplemented with DAP in the anaerobic chamber. Conjugations were performed by plating a 150 µL mixture of BHI plus DAP containing 1 OD_{600}*mL of donor and 1 OD_{600}*mL of recipient community onto filter-topped BHI agar plates, followed by 12 hr incubation at 37°C. Lawns were scraped from filters into 10 mL BHI medium, gently inverted to suspend cells, and 1 OD_{600}*mL from each cell resuspension was used to inoculate both liquid and solid selective medium for outgrowth and enrichment. Selection was facilitated by supplementation of liquid and solid BHI medium with 400 µg mL^{-1} carbenicillin, 400 µg mL^{-1} streptomycin, and 400 µg mL^{-1} spectinomycin. Liquid inoculations in 5 mL BHI were carried out in 24 deep well plates in a stationary incubator at 37°C and sub-cultured using a 100x volumetric dilution into 5 mL fresh medium with antibiotics after 24 hr. After an additional 24 hr incubation at 37°C, subcultures were harvested by centrifugation at 4,000*g* for 10 min and pellets were frozen at -80°C for downstream shotgun metagenomic sequencing and ET-Seq along with pre-conjugation infant gut community samples. Fractional abundance was calculated using the ET-Seq light metagenomics pipeline as described above for pre-conjugation infant gut cultures and post-selection VcDART-targeted cultures. Colonies were picked from selective BHI agar and assayed by colony PCR amplification of the targeted junctions using two primer pairs, the first to detect insertions in the major orientation and the second to detect insertions in the minor orientation, where each primer pair consisted of a genome-specific forward primer and DART-specific reverse primer. Sanger sequencing of amplicons from PCR-positive colonies was used to identify insertion location and barcode, and insertion distances and orientations relative to target sites were plotted for all non-redundant mutants, which were determined as the set of mutants with unique barcode and insertion location combinations (Extended Data Fig. 10).

Statistics and reproducibility

All transformations (natural transformation, conjugation, electroporation) and subsequent analyses were performed for three independent replicates.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Summary data for genomes, plasmids, and oligonucleotides used in this study can be found in Supplementary tables 1 and 3-5. Sequence data for all genomes assembled as part of this study and newly constructed plasmids are in submission to NCBI with accession numbers pending. All genomes and plasmids used in the project will also be made available on ggKbase (https://ggkbase.berkeley.edu/). Raw count data for all experiments including both metagenome and ET-seq information is available at https://github.com/SDmetagenomics/ETsuite/tree/master/manuscript_data.

Code availability

Custom R scripts for ET-Seq analysis and code used in the construction of figures are available at https://github.com/SDmetagenomics/ETsuite.

Acknowledgments

We thank Morgan N. Price for data analysis input, Patrick Pausch for experimental advice, Shana L. McDevitt, Eileen Wagner, and Hitomi Asahara for help with sequencing, and Trent R. Northen for directional advice. Funding was provided by m-CAFEs Microbial Community Analysis & Functional Evaluation in Soils, (m-CAFEs@lbl.gov) a project led by Lawrence Berkeley National Laboratory supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231. Support was also provided by the Innovative Genomics Institute at UC Berkeley. B.E.R. and B.F.C. are supported by the National Institute of General Medical Sciences of the National Institute of Health under award numbers F32GM134694 and F32GM131654. Schematics were created with BioRender.com.

Contributions

B.E.R., S.D., B.F.C, R.B., A.M.D., J.F.B, and J.A.D. conceived the work and designed the experiments. B.E.R., B.F.C., A.L.B., C.H., M.X., Z.Z., D.C.S., K.T., T.K.O., N.K, and R.R. conducted the molecular biology included. S.D., A.C.-C., Y.C.L., H.S., C.H., R.S. and S.J.S. developed the bioinformatic analysis. B.E.R., S.D., B.F.C., Y.C.L., R.B., A.M.D., J.F.B., and J.A.D. analyzed and interpreted the data.

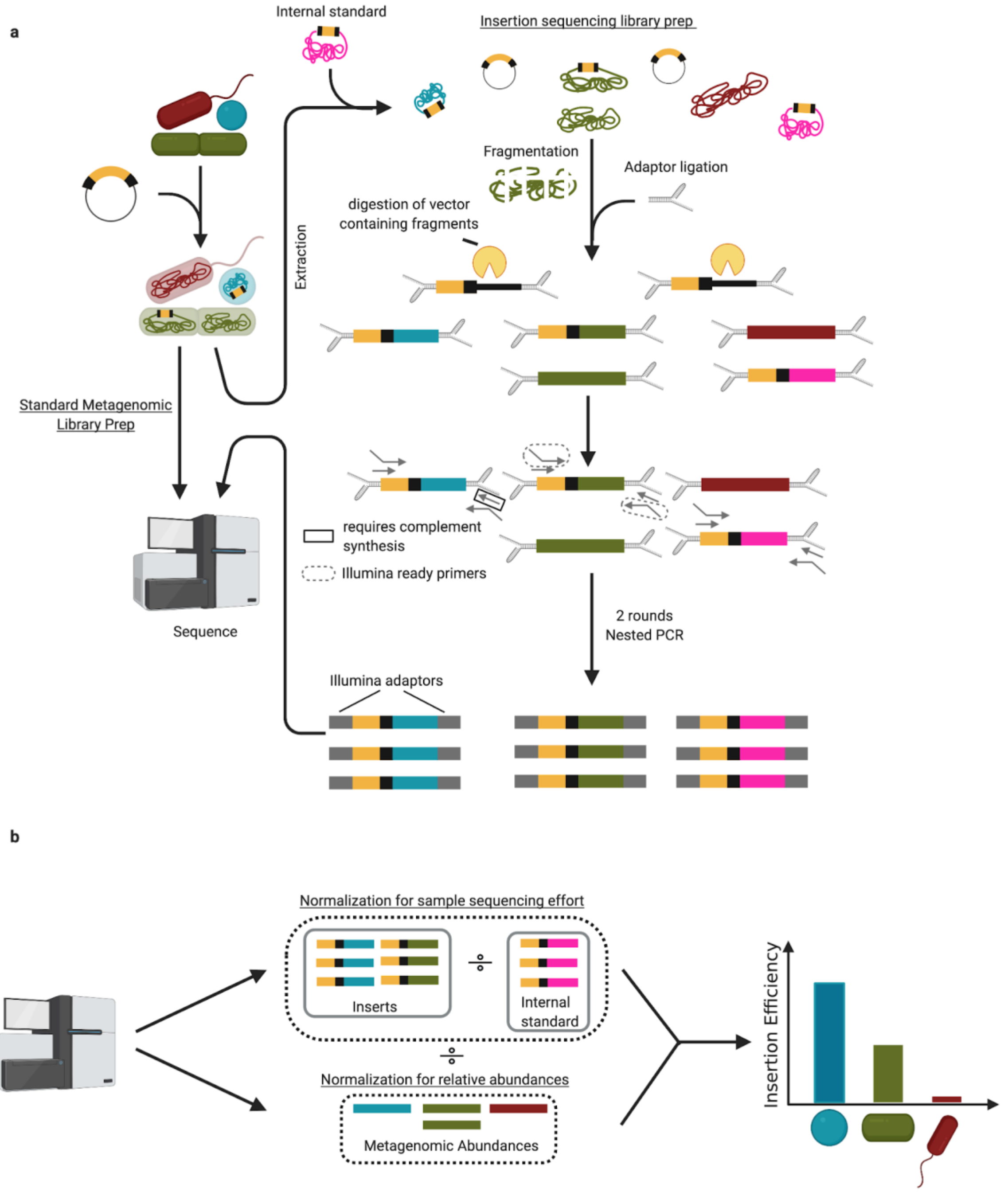
Competing Interests

The Regents of the University of California have patents pending related to this work on which B.E.R., S.D., B.F.C., A.M.D., J.F.B., and J.A.D. are inventors. J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine, Intellia Therapeutics, Scribe Therapeutics and Mammoth Biosciences, a scientific advisory board member of Caribou Biosciences, Intellia Therapeutics, eFFECTOR Therapeutics, Scribe Therapeutics, Synthego, Mammoth Biosciences and Inari, and is a Director at Johnson & Johnson and has sponsored research projects by Biogen, Roche and Pfizer. J.F.B. is a founder of Metagenomi. R.B. is a shareholder of Caribou Biosciences, Intellia Therapeutics, Locus Biosciences, Inari, TreeCo, and Ancilia Biosciences.

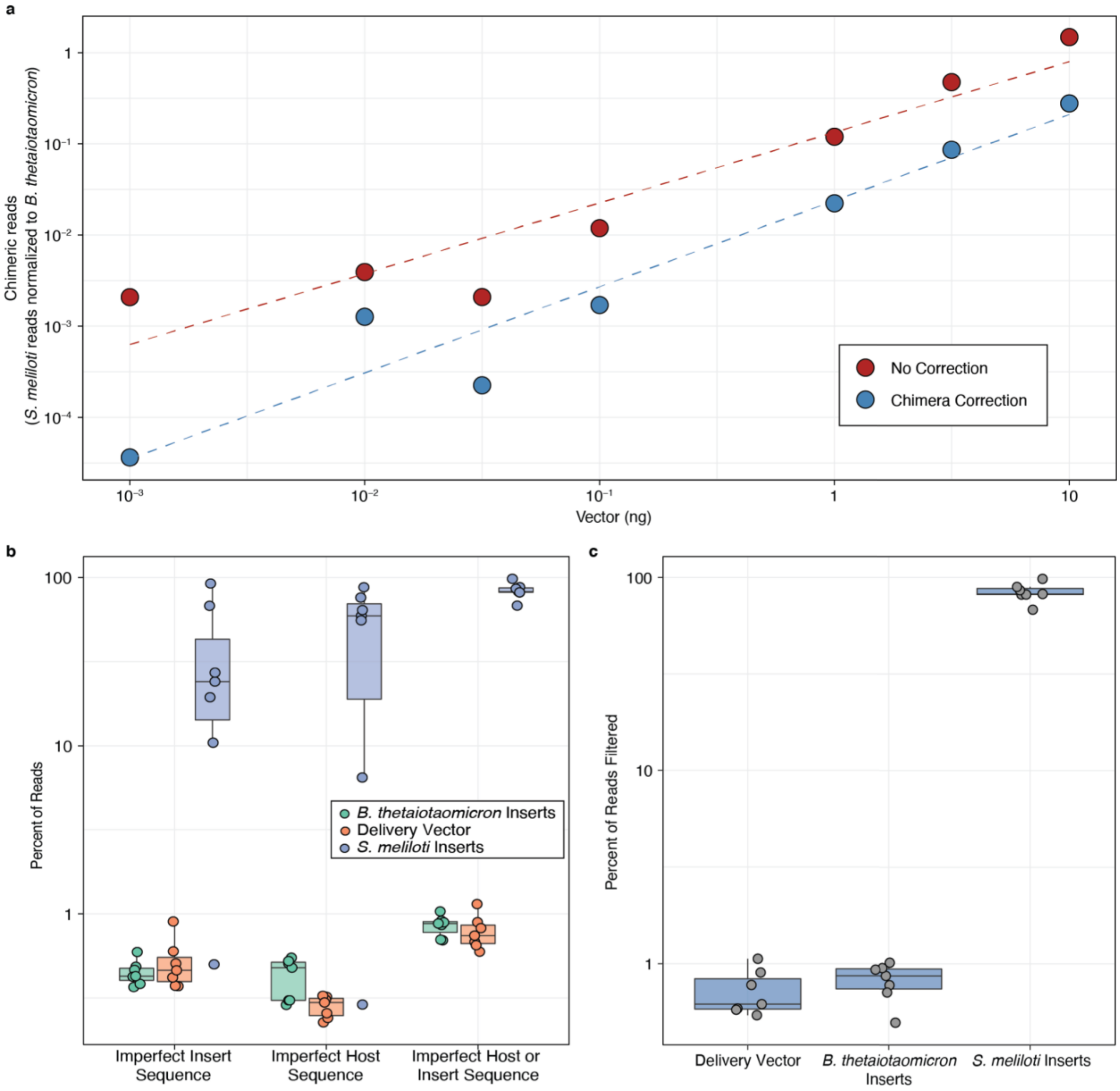
Additional Information

Correspondence and request for materials should be addressed to J.A.D. and J.F.B.

Extended Data

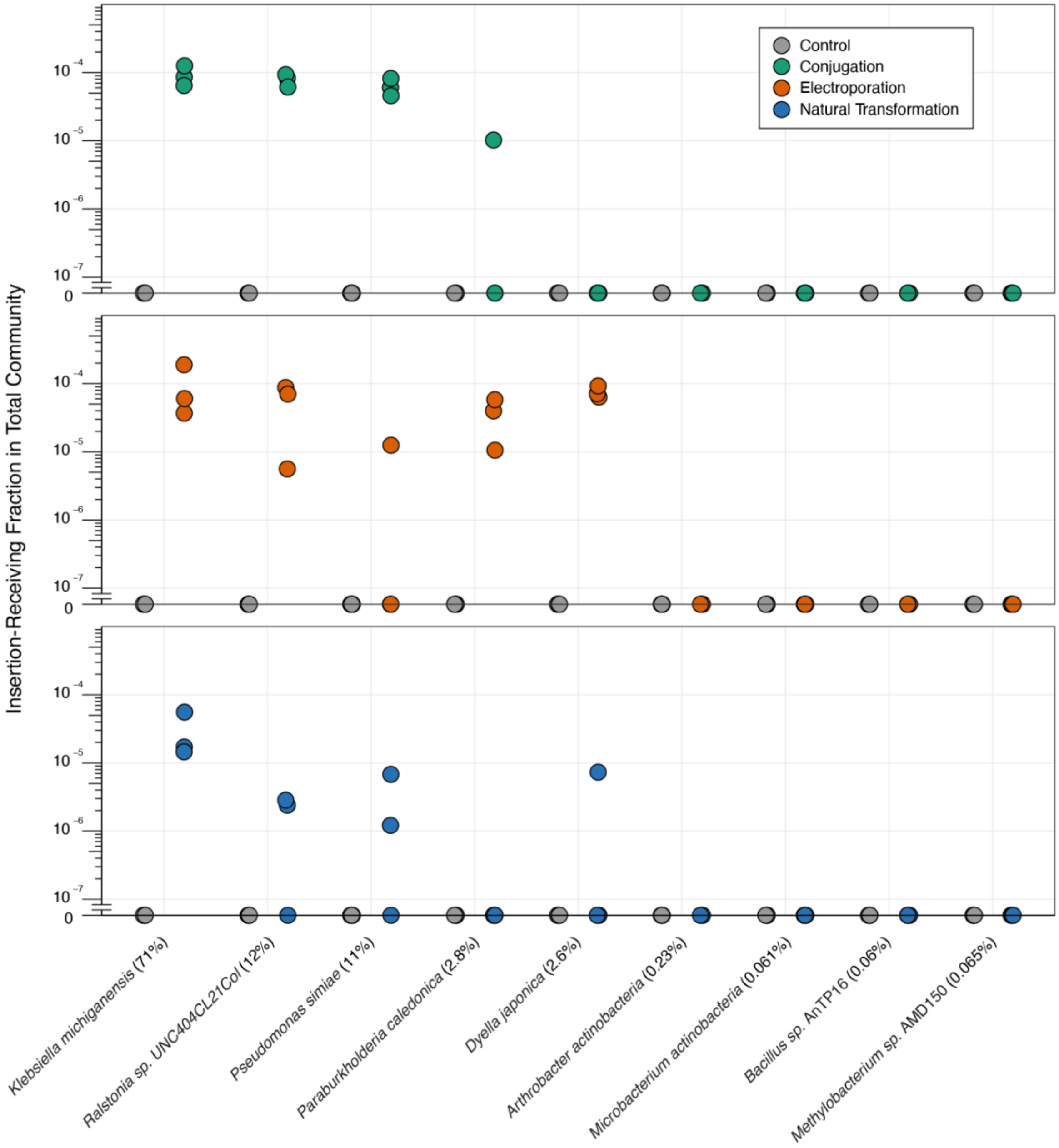

a
Internal standard
Insertion sequencing library prep
Extraction
Fragmentation
Adaptor ligation
digestion of vector containing fragments
Standard Metagenomic Library Prep
Sequence
requires complement synthesis
Illumina ready primers
2 rounds Nested PCR
Illumina adaptors
b
Normalization for sample sequencing effort
Inserts
÷
Internal standard
÷
Normalization for relative abundances
Metagenomic Abundances
Insertion Efficiency

Extended Data Fig. 1 | Library preparation and data normalization for ET-Seq. a, ET-Seq requires low-coverage metagenomic sequencing and customized insertion sequencing. Insertion sequencing relies on custom splinkerette adaptors, which minimize non-specific amplification, a digestion step for degradation of delivery vector containing fragments, and nested PCR to enrich for fragments containing insertions with high specificity. The second round of nested PCR adds unique dual index adaptors for Illumina sequencing. **b**, This insertion sequencing data is first normalized by the reads to internal standard DNA which is added equally to all samples and serves to correct for variation in reads produced per sample. Secondly, it is normalized by the relative metagenomic abundances of the community members.

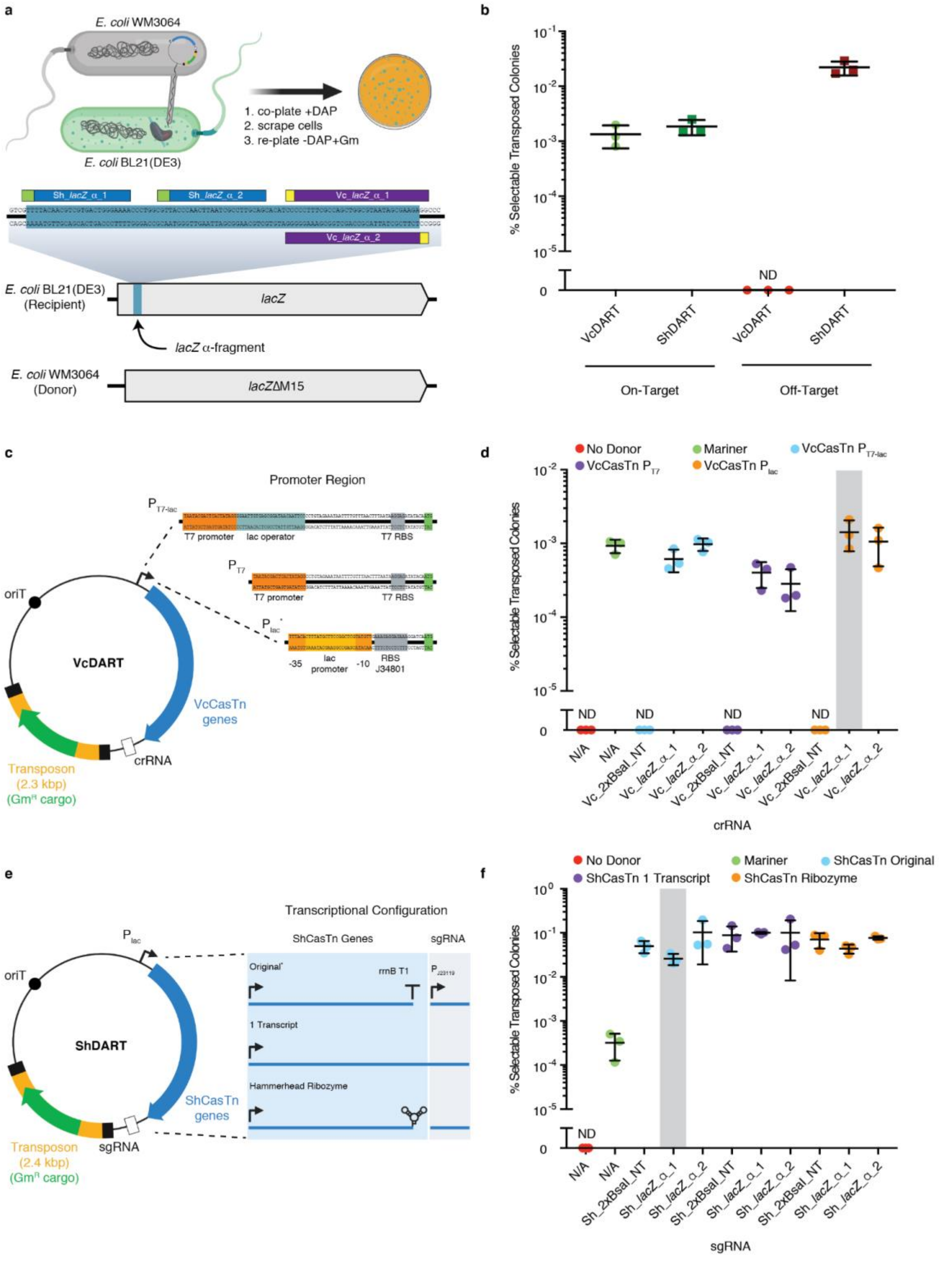


Extended Data Fig. 2 | Measurement and correction of chimeric reads. a, The response of chimeric reads, measured as total normalized read counts to insertions into wildtype *S. meliloti* DNA spiked-in before library preparation, to increasing quantities of donor vector. Plot is log10 scaled on the x and y-axis for readability. Dashed lines indicate log-log linear fit to data ($R^2_{\text{No Correction}}$ = 0.86, n = 7 biological replicates;

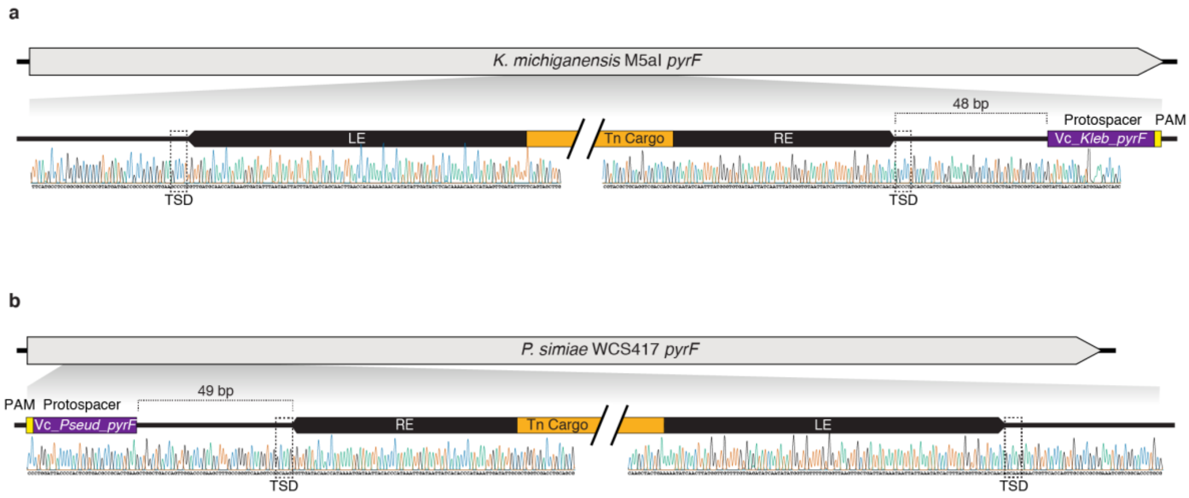
$R^2_{Correction}$ = 0.92, n = 7 biological replicates) **b,** Frequency of read properties (imperfect insert sequence = single difference in last 5 bp of transposon right end from expected sequence; imperfect host sequence = mismatch in first 3 bp of genomic sequence at transposon genome junction when aligned to host genome) identified as strongly associated with *S. meliloti* insertions, in which all reads are expected to be chimeric, used as markers for filtering chimeric reads. Box plots indicate median and bound 1st and 3rd quartile, whiskers indicate max/min values (n = 7 biological replicates). Plot is log10 scaled on the y-axis for readability. **c,** Fraction of insertion mapping reads filtered out of each dataset, for each organism/vector (n = 7 biological replicates) following chimera filtering. Box plots indicate median and bound 1st and 3rd quartile, whiskers indicate max/min values. Plot is log10 scaled on the y-axis for readability.



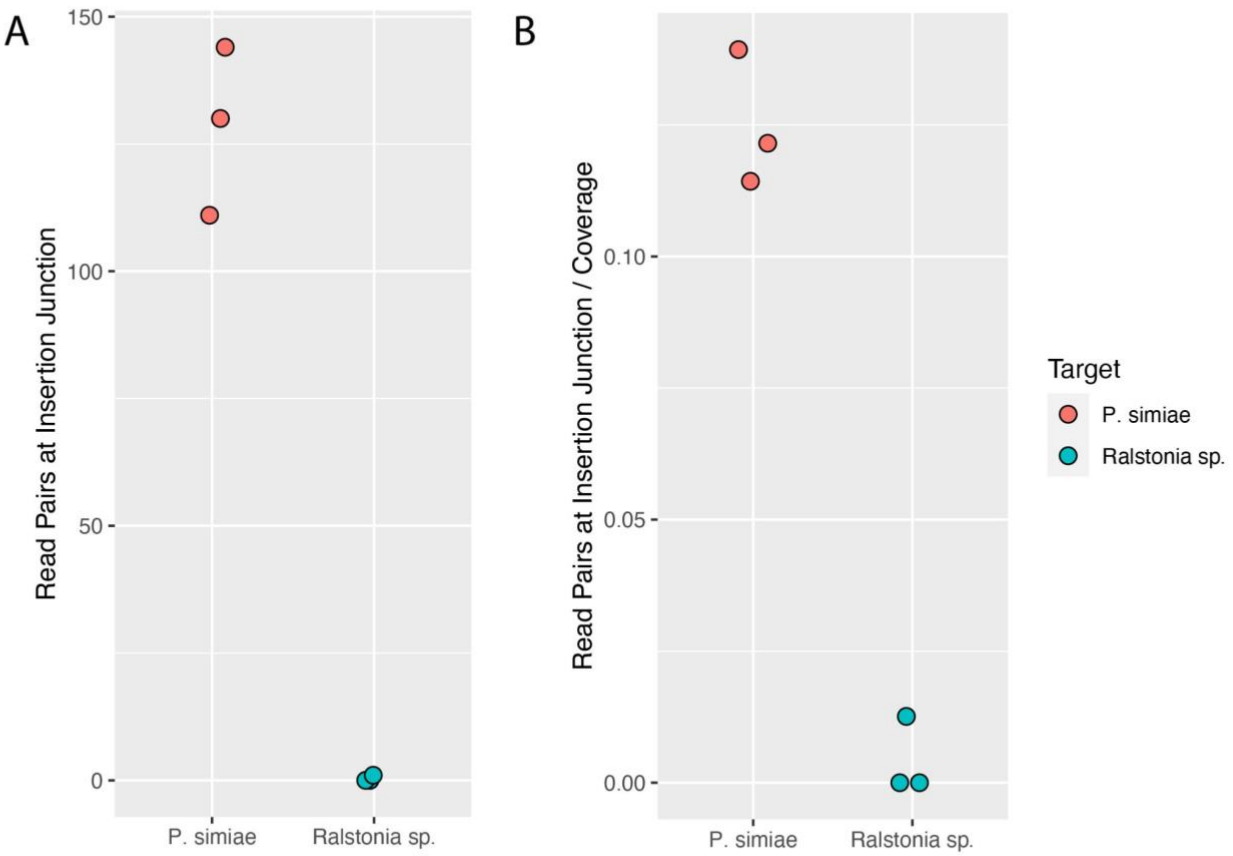
Extended Data Fig. 3 | ET-Seq determined insertion efficiencies for all nine consortium members as a fraction of the entire community. ET-Seq determined insertion efficiencies for conjugation, electroporation, and natural transformation on the synthetic soil community (n = 3 biological replicates). The values shown are the estimated fraction a constituent species's transformed cells make of the total community population. Control samples received no exogenous DNA. Average relative abundance across all samples is indicated in parentheses (n = 18 independent samples).



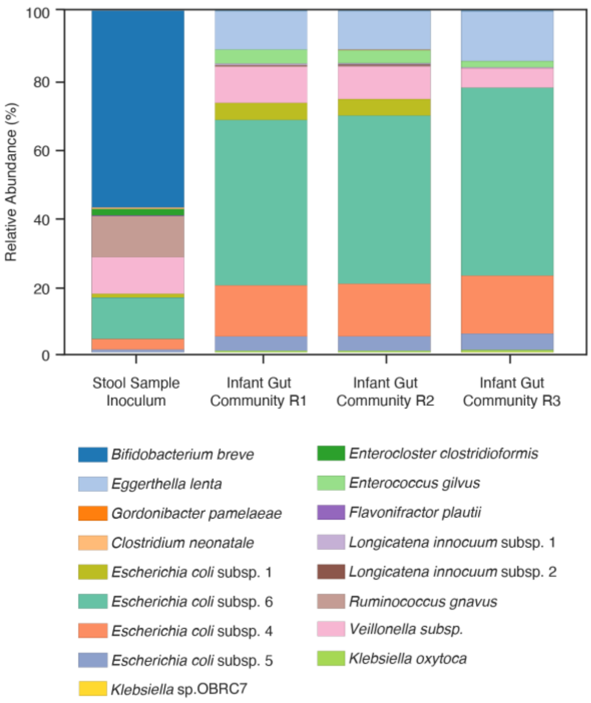
Extended Data Fig. 4 | Benchmarking DART vectors. a, *E. coli* WM3064 to *E. coli* BL21(DE3) conjugation, transposition, and selection schematic (top) and guide RNAs targeting the *lacZ* α-fragment of recipient BL21(DE3), which is absent from donor WM3064 (bottom). **b,d,f,** Percent selectable transposed colonies is calculated as the number of colonies obtained with gentamycin selection divided by total viable colonies in absence of selection. **b,** Insertion-receiving colonies divided into on- and off-targeted. This was calculated by multiplying % selectable colonies for representative guides in **d** and **f** (highlighted by grey bars) by the on- or off-target rates (shown in Fig. 2b). **c**, Transposition with VcDART was tested using three promoters. The variant using the P_{lac} promoter, harvested from pHelper_ShCAST_sgRNA[16], was also used for Fig. 2-5 and Extended Data Fig. 4b, 5, 6, and 8. **d,** Efficiencies of VcDART using various promoters. **e,** Transposition with ShDART was tested with three transcriptional configurations, all using P_{lac}[16]. The configuration used for characterization of ShCasTn originally[16] was also used for Fig. 2 and Extended Data Fig. 4b. **f,** Efficiencies of ShDART using various promoters. **b, d, f,** Crossbar indicates mean and error bars indicate one standard deviation from the mean (n = 3 biological replicates). Guide RNAs ending in "NT" are non-targeting negative control samples.



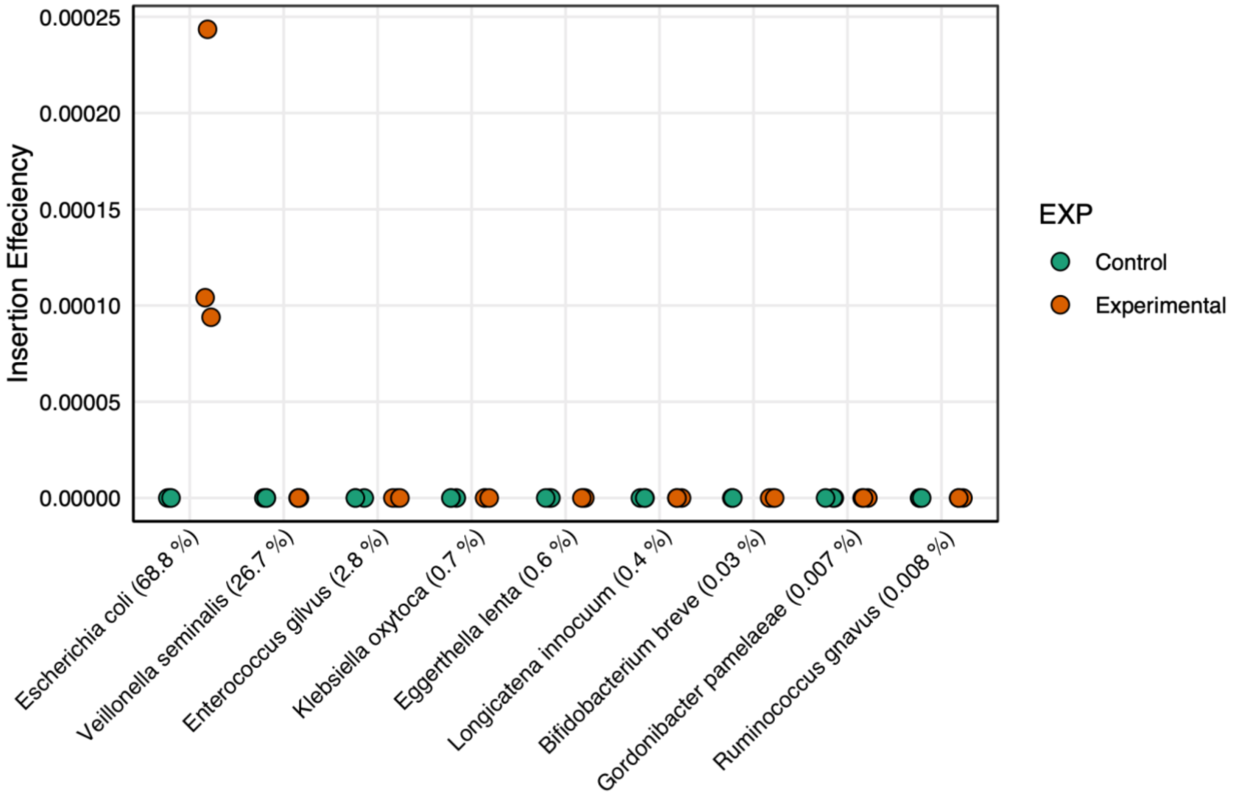
Extended Data Fig. 5 | Sanger sequencing of VcDART mutants from the synthetic soil microbial community. a, Representative Sanger sequencing chromatogram of PCR product spanning transposon insertion site at targeted *pyrF* locus in *K. michiganensis* and **b,** in *P. simiae* mutant colonies following VcDART-mediated transposon integration and selection. Target-site duplications (TSD) are indicated with dashed boxes.



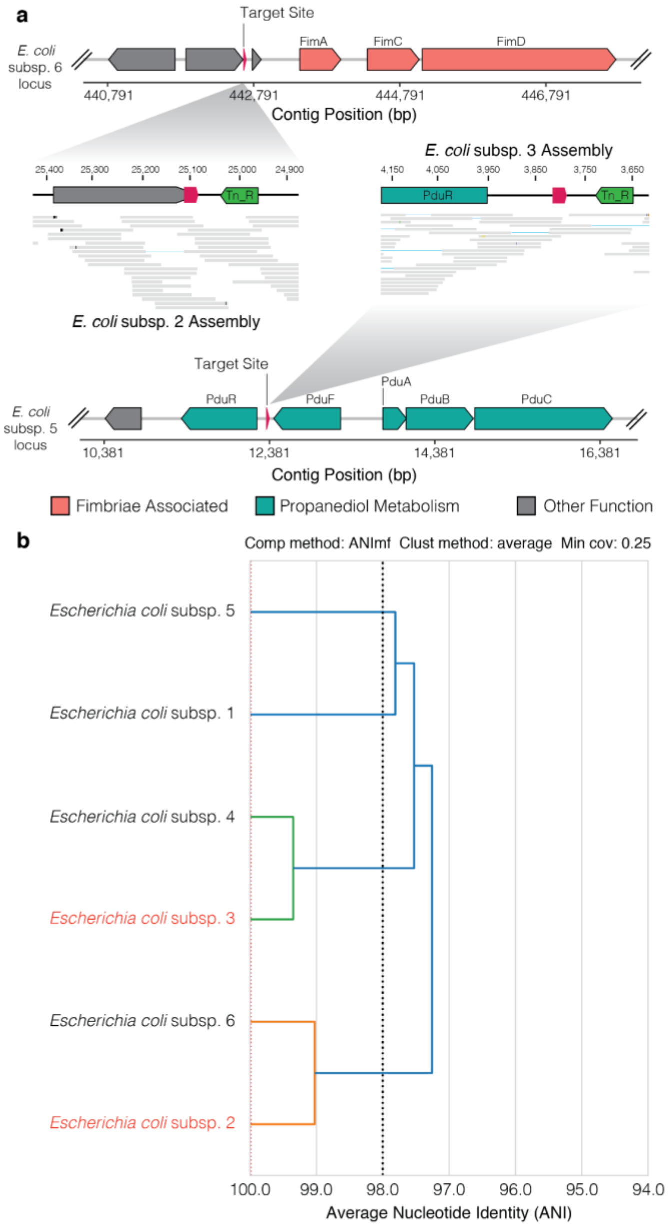
Extended Data Fig. 6 | Insertion counts in *Ralstonia sp.* after metabolic enrichment for *P. simiae*. **a,** Raw number of paired end reads in shotgun sequencing analysis detected as spanning a transposon-genome junction for the *P. simiae* and *Ralstonia sp*. genomes in each of three replicate enrichment samples. **b,** Number of paired end reads detected normalized to the coverage of each genome within each respective sample. The mean number of inserts normalized to coverage were compared between *P. simiae* and *Ralstonia sp*. ($\text{Mean}_{\text{Psim}}$ = 0.1250 ; Mean_{Ral} = 0.0042) and were significantly different (P-value = 0.00058; two-sample t-test).



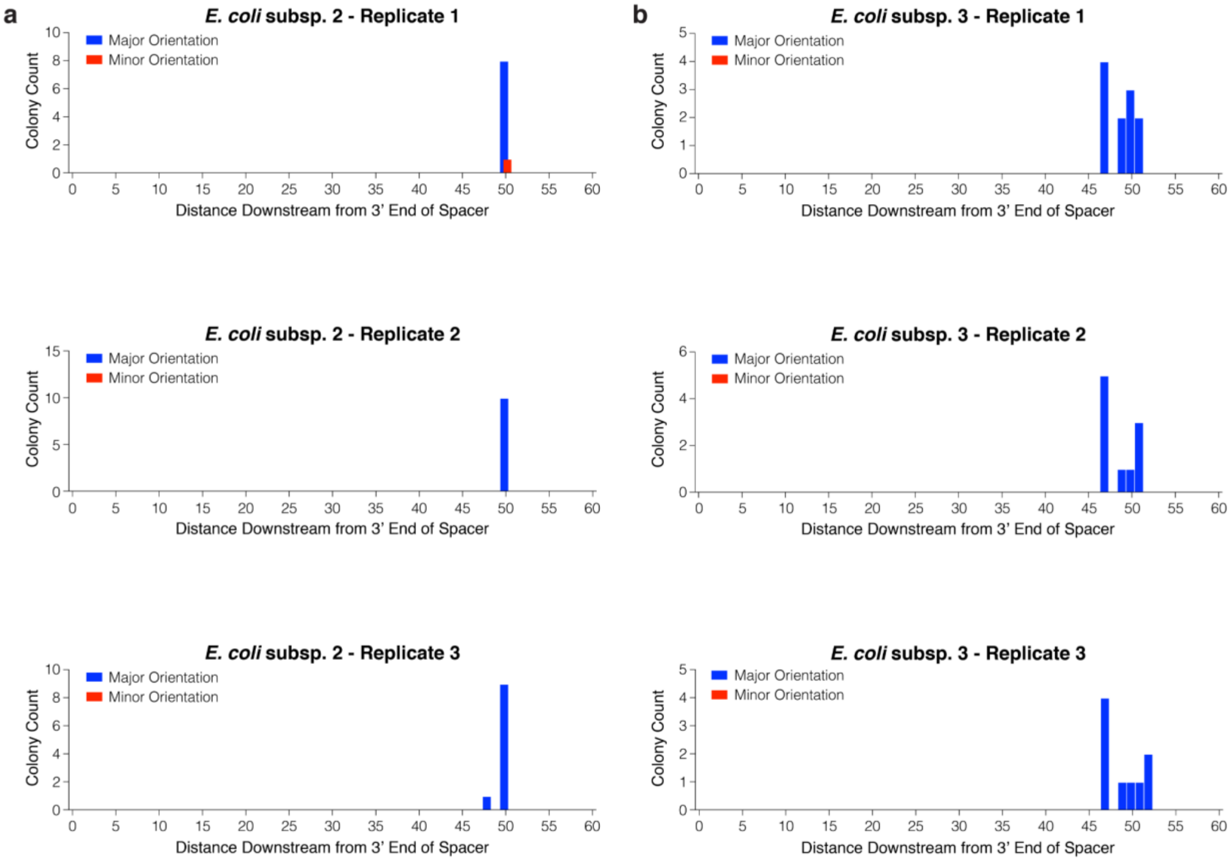
Extended Data Fig. 7 | Relative abundance of stool sample inoculum and infant gut community used for VcDART editing. The gut microbiome compositions were obtained by read mapping to 1005 reference genomes from Lou et al. 2021. Bar height represents normalized subspecies relative abundance, and bars are colored by strain.



Extended Data Fig. 8 | ET-Seq determined insertion efficiency for the infant gut community. Insertion efficiency as quantified by ET-Seq for nine microbial species determined to be present by metagenomic sequencing. Experimental samples were conjugated with a donor containing the unguided mariner transposon (pHLL250; n = 3 biological replicates). Control samples did not receive the donor (n = 3 biological replicates). Percentages next to species names indicate their mean relative fraction in the infant gut community, averaged across the 6 biological replicate experiments performed.



Extended Data Fig. 9: Target site locus and strain comparisons for selective enrichment from infant gut community. a, Clinically relevant gene clusters targeted by VcDART for selective enrichment included a locus associated with fimbriae biosynthesis (top) and a propanediol utilization gene cluster (bottom). Insets show mapped reads to these loci in *E. coli* subsp. 2 and subsp. 3, which were assembled from enrichment culture shotgun sequencing data. The right end of the VcDART transposon cargo was assembled (green), is bridged to the genome, and is supported by paired end read mapping. VcDART target sites (protospacer) are indicated in dark red. **b,** Dendrogram displaying average nucleotide identity differences between all *E. coli* genomes analyzed as part of the infant gut community. Strains in black were genomes originally recovered from metagenomic assembly in Lou, et al. 2021. Strains in red were assembled out of enrichment cultures in this study.



Extended Data Fig. 10 | Location of VcDART transposon insertions in isolated *E. coli* mutant colonies following infant gut community editing. a, Insertion orientations and locations relative to target site were determined by locus-specific PCR and Sanger sequencing on colonies picked from selective solid medium after editing the infant gut community with VcDART guided by the fimbriae associated locus-targeting guide RNA and **b,** the propanediol metabolism locus-targeting guide RNA (n = 3 biological replicates).