New logic facilitated identification of novel phage recombinase function units

Yizhao Chang¹, Qian Wang², Tianyuan Su¹, Qingsheng Qi^{1,3} *

¹ State Key Laboratory of Microbial Technology, Shandong University, No. 72 Binhai Road, Qingdao 266237, P.R. China

² National Glycoengineering Center, Shandong University, Jinan, Shandong, China

³ CAS Key Lab of Biobased Materials, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong, China

* Address correspondence to Qingsheng Qi, qiqingsheng@sdu.edu.cn.

Abstract

Phage recombinase function unit (PRFU) contains one single strand annealing protein (SSAP) and one 5'-3' exonuclease (EXO). Well-known PRFU such as lambda-Red or Rac RecET has been proved to be useful tool in the recombineering of *Escherichia coli* but had limited efficiency in distant species. Therefore, PRFUs from closely related species were often searched and developed as specific genetic tools. However, there are still many bacteria with few PRFU identified. Here, we hypothesised that the occurrence of PRFU might be related to specific strains. Based on this assumption, we managed to identify 59 unique sets of PRFUs in over 23 species of the genus *Corynebacterium*, a taxonomic group that were reported to be lacking in such system. The identified PRFUs were then classified to show their similarities and relation to species based on both sequential and structural information. Representative PRFUs were verified to be highly effective in mediating recombineering using ssDNA or dsDNA substrates. Analysis of the functional PRFUs indicated that they share similar local genome contexts, which might suggest their common origin. Our findings reveal the relation of PRFUs, species and strains and provide novel candidate genetic manipulation tools in genus *Corynebacterium*.

Author Summary

Genus *Corynebacterium* is a taxonomic group with highly diversified species that are of medical, veterinary, or biotechnological relevance. However, being lack of genetic tools limits the study of these species. Our new verified phage recombinase function units (PRFUs) in this genus that

are highly effective in mediating recombineering will provide new candidate genetic tools in this genus. Besides, reference genomes were often used for genome study and data mining. While for the mobile genetic elements, their transferable characteristic may affect their occurrence among genomes. Here we provide evidence for this idea by showing that the occurrence of PRFU is strain related but not species related. This may be helpful for the functional genomics study of mobile genetic elements among genomes.

Introduction

Phage recombinase function unit (PRFU) has been proved to be powerful genetic tool in the recombineering of some bacteria [1-4]. This system mainly contains one single strand annealing protein (SSAP) and one 5'-3' exonuclease (EXO) [5]. Expression of the SSAP alone can mediate recombineering with single-strand DNA (ssDNA) substrates, while expression of the whole system could use double-strand DNA (dsDNA) as substrates for genome fragment editing[1, 2].

Well-known PRFU such as lambda-Red or Rac RecET has been used for successful recombineering in *Escherichia coli*. The Red system is phage lambda derived, while the RecET system is from the defective prophage Rac of *Escherichia coli* [1]. Lambda-Red contains three adjacent genes *gam*, *bet* and *exo*, which encode a host nuclease inhibiting protein Gam, a SSAP Beta and an EXO accordingly [6]. In *E. coli*, it is possible to mediate recombineering using dsDNA fragment with homology arm as short as 35 base pairs through the expression of Red-EXO, Red-Beta and Red-Gam [7]. RecET, on the other hand, contains only one SSAP RecT and one

exonuclease RecE, which has been shown to be highly efficient in linear-linear homologous recombination than lambda-Red [2, 8]. Both lambda-Red and Rac RecET function by specifically interacting with their respective partners, except for the Red-Gam, which is not necessary for dsDNA recombination but can improve the efficiency [1, 9].

Previous studies have showed that PRFU functioned well in related species but performed poorly in distant species, presumably because the exonuclease is more species-specific and could not function well heterogeneously [1, 6]. Thus, the PRFUs of species within the same genus were often searched and developed for specific genetic tool [1, 4, 6, 7]. Typically, analogies of Beta or RecT were often searched, and EXOs besides these SSAPs were also compared [10, 11]. However, several species have been reported to have no such system in this way, such as those in genus *Corynebacterium* [12, 13].

Genus *Corynebacterium* is a taxonomic group with highly diversified species that are of medical, veterinary, or biotechnological relevance[14]. The medical and industrial value encourage the development of genetic manipulation tools for the species of this taxon [13, 15-17]. Up to now, only four species in this genus have been reported to contain PRFU [13, 18]. Even in the well-studied *C. glutamicum*, no PRFU has been identified yet [13].

Here, we proved the idea that the occurrence of PRFU might be related to specific strain by identifying the PRFUs in genus *Corynebacterium*. The identified PRFUs were classified to reveal their relation to species. Typical PRFUs were experimentally verified in mediating recombineering using ssDNA or linear dsDNA substrates. Functional PRFUs were found to share conserved local genome contexts, which might suggest their common origin. These findings

provide novel candidate genetic manipulation tools in genus *Corynebacterium* as well as new insight for functional genomics study.

Results

Identification of novel PRFU using a database to database method

For most of the studies relating to PRFU, reference genomes were used for bioinformatics analysis [1, 12, 18, 19]. However, as previous verified PRFUs are mainly from prophages or defective prophages, we believe their occurrence may be related to specific bacterial strains. So, all the bacteria genomes as well as the phage genomes of genus *Corynebacterium* in the NCBI genome database were used for study.

As PRFU is mainly composed of one SSAP and one EXO, which usually appear next to each other, we studied the existence of both among genomes to improve prediction accuracy. A database to database method was developed for the searching process. The conserved domain database of SSAPs and EXOs (SSAPdb and EXOdb) were first created using sequences from NCBI conserved domain database. The protein sequences of each genome were then used as queries for PSI-BLAST in the SSAPdb and EXOdb. Then consensus sequences were got through multiple sequence alignment after adding candidate SSAPs or EXOs found to these databases. Then new searches were done until no new SSAP or EXO were found (Fig 1). After iterating for 3 rounds, we identified PRFUs in 84 of the 429 genomes studied, including both classified species and unclassified genomes (S1a-b Fig). As a result, the SSAPs were found in 84 genomes and EXOs

were found in 81 genomes. Among them, 54 of the 59 unique SSAPs and 56 of the 57 unique EXOs that were previously annotated as hypothetical protein were identified.

When searching for PRFU among all the plasmids of genus *Corynebacterium*, only one SSAP was found on the plasmid of *C. glutamicum* R strain. This may explain why no PRFU was found previously in *C. glutamicum* when searching its reference genome [12, 13]. In addition to the bacteria genomes, the verified phage genomes of genus *Corynebacterium* were also analyzed but no PRFU was found.

To prove the origin of these PRFUs, the prophages were predicted using an integrated search and annotation tool PHASTER [20]. It was found that among the 84 genomes containing PRFUs, 52 of them have PRFU in the prophage region, while 27 of them have similar proteins coded next to PRFU as those 52 genomes (S1c Fig).

Analysis of the genomes showed that these SSAPs and EXOs form 59 sets of unique PRFUs in over 23 species (Table 1). Most of these species were reported to have no PRFU system or only have SSAP previously [12, 13, 21]. Different from other species, *C. argenteorotense* and *C. glutamicum* both separately have one set of incomplete PRFU with only SSAP but no EXO. Besides, not all of the genomes sequenced in these 23 species have PRFU. To test whether this is caused by the incomplete genome sequence data or it just reflects the relationship of PRFUs, genomes and species, the assemble level of the genomes were analyzed. The result shows that the existence of PRFU has no difference between the completely assembled genomes and the not completed ones (S1 Table, t test, $p = 0.1314$), which indicates that the PRFU occurrence is genome (strain) related but not species related.

Classification of the PRFUs and their components

To study the relation of these PRFUs, their component SSAPs and EXOs were classified into 5 types and 3 types separately based on both protein sequence similarity and secondary structure similarity (Figs 2). Each type is conserved in sequence as well as second structure and lies within one branch of the phylogenetic trees. Some identified SSAPs or EXOs from other bacteria were also incorporated.

For SSAPs, the residues in helices are conserved in the same type, while the residues in strands tend to vary from sequence to sequence. Multiple sequence alignment result showed that SSAP-2 shares much similar to Beta, whereas SSAPs in SSAP-5 were found to have sequence similarity with RecT. A conserved β - β - β - α fold structure was observed in the predicted structure of SSAP-1, SSAP-3 and SSAP-4, while a deformed β - β - β - β - α fold structure was observed in SSAP-2 and SSAP-5 (Fig 2c). The same β - β - β - α fold has been observed in human Rad52, which was found to be composed of highly conserved amino acid residues that are responsible for ssDNA and dsDNA binding [22]. The structure of the N-terminal of human Rad52 (Rad52₁₋₂₁₂) was shown in Fig 2d as an indicant. Actually, a deformed β - β - α - β - β - α fold of Rad52₁₋₂₁₂ was observed in all SSAP types except SSAP-1. Besides, similar to the Rad52₁₋₂₁₂, both the C-terminal and the N-terminal of the SSAPs are rich in helices. However, compared to the conserved N-terminal, the residues in the C-terminal region of different SSAP types are irregular. This might be accordingly with the fact that the conserved N-terminal of SSAP is responsible for homologous pairing, while the C-terminal interacts with EXO or host factors [23-26].

Unlike SSAPs, the second structure of EXOs varied a lot from type to type (Fig 2e). The residues in the N-terminal half of EXOs are conserved across types, while the residues in the C-terminal half vary a lot even in the same type. The residue Leu-91 and residue Phe-94 of λ Exo have been found to be important for forming the λ Exo-Beta complex [26]. These residues lie within a helix of the α - α - β - β fold of the λ Exo (Fig 2f), which was also observed in EXO-3.

These SSAPs and EXOs form eight types of PRFU in total, which is less than their theoretical combination (Fig 2g). Typically, SSAP-1 only appear with EXO-1 and forms PRFU-I, while SSAP-5 and EXO-2 only appear with each other and form PRFU-VII. Interestingly, all the species that harbour more than one type of PRFU are pathogenic (S2 Fig), while not all pathogenic species have multi-type PRFUs.

Verification of the function of PRFUs in vivo

Previous study indicated that the expression of SSAP alone could promote ssDNA mediated recombineering, while expression of the full PRFU would facilitate genetic manipulation of genome with dsDNA [1, 6, 7, 27]. To verify the function of PRFUs identified, both oligo mediated recombination and dsDNA mediated recombination experiments were conducted in a previously constructed strain C.g-kan(-) [28].

To verify the function of SSAPs, eight SSAPs were expressed on the shuttle expression vector PXMJ19 (Fig 3a). The SSAPs were chosen based on the species evolutionary relationship (inferred from the phylogenetic tree of their 16s rRNA, S3 Fig) as well as the PRFU type. An oligo

targeting the defective *kan*(-) was used for ssDNA mediated recombineering to recover the function of *kan* (Fig 3b, [13, 28]). As a result, five of them could promote ssDNA mediated recombination in *C. glutamicum* (Fig 3c-d, Table 2 Sap1-Sap8). While the expression of WP_010191133.1 from *C. aurimucosum* got the highest recombinant number and the highest recombination frequency, the recombination frequency of WP_052659100.1 from *C. diphtheriae* and WP_044027532.1 from *C. glutamicum* was similar to that of the positive control RecT from *E. coli*. No or extremely low level of recombination phenomenon was observed in strains expressing WP_060941986.1 from *C. Kroppenstedtii*, WP_018296493.1 from *C. lubricantis*, and WP_049159720.1 from *C. striatum*.

To verify the recombineering ability of the full PRFUs whose SSAPs were proved to be functional above, their component EXOs were expressed after the corresponding SSAPs (Fig 4a). Selection marker interspaced by 800bp homology arm targeting *recA* was used as dsDNA substrate (Fig 4b). Though no complete PRFU was found in *C. glutamicum* R strain, the gene next to its SSAP coding gene was also tested. It was found that both PRFU from *C. aurimucosum* and RecET could promote highly efficient recombination in vivo with the recombination frequency at 10^{-2} magnitude (Table 3). The PRFU from *C. halotolerans* was also proved to be functional, while was an order of magnitude less efficient. No or extremely low recombination frequency was observed in the PRFUs from *C. glutamicum* and *C. diphtheriae*. The experimental evidences were shown in Fig 4c-e.

Discussion

Although genome sequence grew rapidly these years, few PRFUs have been identified. This and the phage origin of verified PRFUs led us to investigate all the genomes sequenced in genus *Corynebacterium*. As each genome represent a strain record, the occurrence of PRFUs in both reference and non-reference genomes and the irregular distribution pattern of PRFUs among genomes within the same species verified the idea that the existence of PRFUs is strain related but not species related. This indicates that the prediction of PRFU using reference genomes might lead to error conclusion about their existence in the species.

For most of the time, sequence homology-based method was used for the prediction of protein function [29]. When predicting SSAP, a single query such as RecT or Red-Beta was used to start the search for new homologies[13, 19]. To improve the prediction accuracy, the protein sequences of genomes were searched for several rounds in the conserved domain database of SSAPs or EXOs in this study. This enabled us to select results that were identical to several records in the conserved domain database. Besides, by doing iteration, distant-related SSAPs or EXOs were also identified.

To test the function of PRFUs found in this study, we verified several of them with both ssDNA and linear dsDNA mediated recombineering heterogeneously in *C. glutamicum*. Several tested PRFUs were proved to be functional in promoting ssDNA or linear dsDNA mediated recombination. Actually, some PRFUs tested showed relatively high ability in conducting recombination as previously reported tools in other bacteria [1, 6, 7]. As these PRFUs can function in *C. glutamicum*, they are believed to function even better in their hosts for their

specific molecular interactions with the host factors [1]. This suggested their great potential for recombineering in genus *Corynebacterium*.

Four of the five functional SSAPs verified in this study have more closed local genome context relation than others and lie within the same branch of the local genome context tree (Fig 5a). All of the S_{dis} of local genome context to *C. glutamicum* R strain in this branch are smaller than 35 and have a wider range from 0 to 35 compared to others, which have a narrow range of S_{dis} from 35 to 40 (Fig 5b). This suggests that the local genome context conservation might be related to the function of SSAPs. Additional verified SSAPs with S_{dis} smaller than 35 functioning in the recombineering experiment verified this idea (Table 2, Fig 5). Analysis of the components of the local genome contexts in this branch indicates that they are rich in genes whose products may interact with DNA, such as transcriptional regulators and site-specific integrase. Besides, no linear relation between the species 16S rRNA distance and local genome context distance was observed (S4 Fig). These suggest that the local genome context conservation in this branch might be caused by the horizontal gene transfer.

In this study, the database to database searching method facilitated accurate prediction of novel PRFUs in the species that were reported to have no such system. Analysis of the distribution of PRFUs among genomes indicates that their occurrence is related to strains but not species. This may provide new insights for the study of mobile genetic elements. Besides, several PRFUs were proved to be highly effective in mediating recombineering using ssDNA or dsDNA substrates in this study. Due to the limited genetic manipulation tools available in genus

Corynebacterium, the PRFUs verified will also provide new candidate tools in these and related species.

Materials and methods

Analysis of PRFU across genomes

Genomes of genus *Corynebacterium*, including chromosomes, plasmids and phages were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>) genome database. The conserved domain data of SSAP and EXO from the Conserved Domains Database of NCBI (Accession number: cl04285, cl04500, cl01936 and pfam09588) were also downloaded. Then the following process were iterated:

1. Multiple sequence alignments were done to get the consensus of SSAPs and EXOs and construct database SSAPdb as well as EXOdb.
2. Then genomes of genus *Corynebacterium* were searched for potential SSAP and EXO in the SSAPdb and EXOdb separately through PSI-BLAST for one iteration.
3. Potential SSAP and EXO were extracted and added to the SSAPdb or EXOdb accordingly.

This process was iterated for three times until no new potential SSAP or EXO was found. Potential SSAPs were chosen from the result of PSI-BLAST using a cut off E value of 10^{-6} . Potential EXOs were chosen by setting a cut off E value as 10^{-5} .

Prophage was predicted by PHASTER (<http://phaster.ca/>) [20]. The genome location of PRFUs as well as the local genome context were used for the verification of their prophage origin.

Protein sequence and structure analysis

Multiple protein sequence alignment was performed with MUSCLE [30]. The alignment results were then edited manually to get the consensus parts and shown as the default settings in ClustalX [31].

All of the second structures of the alignments were predicated by Jpred4 (<http://www.compbio.dundee.ac.uk/jpred4/index.html>) [32].

Alignment of the local genome context

The adjacent five proteins coded upstream and downstream the SSAP of each genome were extracted to evaluate the local genome context similarity if available. First, the similarity of these proteins was analyzed through PSI-BLAST. Protein pairs with E value less than 10^{-5} were grouped together. The Groups with proteins of similar annotation or the proteins that were not grouped in the previous step but have similar annotation were merged into one group. The proteins that were not grouped yet were each assigned a group separately.

A similarity matrix of these local genome context was then calculated through the similarity score S_{sim} between each pair of them. The similarity score S_{sim} contained two parts, S_{sam} and S_{loc} . S_{sam} was calculated as the sum number of protein pairs from the two genomes that are of same group. S_{loc} was the sum of S_i , which was calculated in the following way: For each protein coded upstream SSAP and downstream SSAP, its position was assigned a site i . Site i has a range from

1 to 11, corresponding to the position from the farthest upstream the position of SSAP to the farthest downstream. If the two proteins of both genome at the site i belongs to the same group, then S_i was assigned 1,2,3,4,5,6,5,4,3,2 or 1 for each i from 1 to 11 accordingly. This can be written as:

$$S_{sim} = \sum_{i=1}^{11} S_i + \sum S_{sam} \quad (1)$$

The distance matrix of these genomes was got from the similarity matrix, which used a score S_{dis} . S_{dis} was calculated as:

$$S_{dis} = 47 - S_{sim} \quad (2)$$

The number 47 was the S_{sim} for a pair of completely same local genome context that had an ideal S_{dis} value of 0. As different site was assigned a different weight, we call these site-weighted scores.

Phylogenetic analysis

The genome phylogenetic tree was constructed using the 16S rRNA of the representative genomes. The protein phylogenetic tree was constructed using the multiple sequence alignment results of SSAPs and EXOs. For the construction of local genome context phylogenetic tree, a distance matrix was used as input. The phylogenetic trees were constructed by MEGA6 using Neighbor-Joining method with the default settings except for the Interior-branch Test value, which is set as 1000 [33].

280

281 **Strain cultivation**

282 The *C. glutamicum* strain C.g-kan(-) was cultured in LB or BHIS medium. For the preparation of
 283 the electrocompetent cells, the strains were first cultured in LB medium overnight at 30°C. The
 284 cultures were then added to 100ml BHIS medium at a start optical density of 0.3 and cultured at
 285 30 °C until the optical density reached 0.9, with IPTG and antibiotics added according to the
 286 circumstance. The culture was then chilled on ice for 30min and centrifuged at 4°C to collect
 287 the cells. The sediments were washed for four times using 30ml 10% (v/v) glycerol and
 288 resuspended in 1ml 10% (v/v) glycerol. Then a volume of 100ul each tube was divided and
 289 stored at -80°C for later experiment. To maintain the plasmids, a final concentration of 100
 290 µg/ml spectinomycin, 15µg/ml kanamycin or 10 µg/ml chloromycetin was added to the
 291 corresponding cultures. For induction purpose, a final concentration of 0.25 mmol/L IPTG was
 292 used [34].

293

294 **Statistics**

295 The ratio of the number of sequence finished genomes with PRFU and sequence finished
 296 genomes as well as the ratio of the number of sequence unfinished genomes with PRFU and
 297 sequence unfinished genomes were used for t-test (S1 Table).

298

Recombineering with PRFUs

The plasmid, ssDNA and dsDNA substrates were transformed into *C. glutamicum* electrocompetent cells by electroporation. Different substrates were pre-chilled and added to the electrocompetent cells according to the circumstances. For recombineering with ssDNA, 2µg oligo was added to the competent cells. For plasmid and dsDNA, 500ng or 1µg substrate was used. The mixture was then used for electroporation at 1.8kv/mm, 200Ω, 25µF. Then 1ml 46°C BHIS medium was instantly added. The mixture was heat shocked at 46 °C for 6min and then cultured for outgrowth at 30 °C, 200rpm up to 1h for strain construction purpose or 30 °C, 150rpm for 5h for the transformation of ssDNA, dsDNA and control plasmid. Cultures were then plated directly after collecting the cells for most of the time. To count the number of viable cells, 0.1ml of 10⁻⁴ dilution was plated on BHIS with chloromycetin added [13, 35].

Fragments were prepared through overlap extension PCR using TransStart FastPfu Fly DNA Polymerase (TransGen, AP231-03) and collected using DNA Gel Extraction Kit (TsingKe, GE101-200). Results were verified by agarose gel electrophoresis and Sanger sequence.

Acknowledgements

We thank Haiying Jin for her help in the recombineering with ssDNA experiment. We also thank Qilong Qin for reviewing the manuscript.

Author contributions

YZ designed and conducted the experiment, prepared and revised the manuscript. QS designed, supervised the study and revised the manuscript. QW revised the manuscript. TY participated the experiment.

References

1. Datta S, Costantino N, Zhou X, Court DL. Identification and analysis of recombineering functions from Gram-negative and Gram-positive bacteria and their phages. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(5):1626-31. Epub 2008/01/31. doi: 10.1073/pnas.0709089105. PubMed PMID: 18230724; PubMed Central PMCID: PMC2234195.
2. Zhang Y, Buchholz F, Muyrers JP, Stewart AF. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat Genet*. 1998;20(2):123-8. Epub 1998/10/15. doi: 10.1038/2417. PubMed PMID: 9771703.
3. Murphy KC. Use of bacteriophage lambda recombination functions to promote gene replacement in *Escherichia coli*. *Journal of bacteriology*. 1998;180(8):2063-71. PubMed PMID: 9555887; PubMed Central PMCID: PMC107131.
4. Yang P, Wang J, Qi Q. Prophage recombinases-mediated genome engineering in *Lactobacillus plantarum*. *Microbial cell factories*. 2015;14:154. Epub 2015/10/07. doi: 10.1186/s12934-015-0344-z. PubMed PMID: 26438232; PubMed Central PMCID: PMC4595204.
5. Vellani TS, Myers RS. Bacteriophage SPP1 Chu is an alkaline exonuclease in the SynExo family of viral two-component recombinases. *Journal of bacteriology*. 2003;185(8):2465-74. Epub 2003/04/03. PubMed PMID: 12670970; PubMed Central PMCID: PMC152610.
6. van Kessel JC, Hatfull GF. Recombineering in *Mycobacterium tuberculosis*. *Nat Methods*. 2007;4(2):147-52. doi: 10.1038/nmeth996. PubMed PMID: 17179933.
7. Yin J, Zhu H, Xia L, Ding X, Hoffmann T, Hoffmann M, et al. A new recombineering system for *Photobacterium* and *Xenorhabdus*. *Nucleic acids research*. 2015;43(6):e36. Epub 2014/12/30. doi: 10.1093/nar/gku1336. PubMed PMID: 25539914; PubMed Central PMCID: PMC4381043.
8. Fu J, Bian X, Hu S, Wang H, Huang F, Seibert PM, et al. Full-length RecE enhances linear-linear homologous recombination and facilitates direct cloning for bioprospecting. *Nature biotechnology*. 2012;30(5):440-6. Epub 2012/05/01. doi: 10.1038/nbt.2183. PubMed PMID: 22544021.
9. Muyrers JP, Zhang Y, Buchholz F, Stewart AF. RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. *Genes & development*. 2000;14(15):1971-82. Epub 2000/08/02. PubMed PMID: 10921910; PubMed Central PMCID: PMC316823.

10. Swingle B, Bao Z, Markel E, Chambers A, Cartinhour S. Recombineering using RecTE from *Pseudomonas syringae*. *Appl Environ Microbiol*. 2010;76(15):4960-8. Epub 2010/06/15. doi: 10.1128/aem.00911-10. PubMed PMID: 20543050; PubMed Central PMCID: PMC2916495.
11. Van Pijkeren JP, Neoh KM, Sirias D, Findley AS, Britton RA. Exploring optimization parameters to increase ssDNA recombineering in *Lactococcus lactis* and *Lactobacillus reuteri*. *Bioengineered*. 2012;3(4):209-17. Epub 2012/07/04. doi: 10.4161/bioe.21049. PubMed PMID: 22750793; PubMed Central PMCID: PMC3476877.
12. Resende BC, Rebelato AB, D'Afonseca V, Santos AR, Stutzman T, Azevedo VA, et al. DNA repair in *Corynebacterium* model. *Gene*. 2011;482(1-2):1-7. doi: 10.1016/j.gene.2011.03.008. PubMed PMID: 21497183.
13. Binder S, Siedler S, Marienhagen J, Bott M, Eggeling L. Recombineering in *Corynebacterium glutamicum* combined with optical nanosensors: a general strategy for fast producer strain generation. *Nucleic acids research*. 2013;41(12):6360-9. doi: 10.1093/nar/gkt312. PubMed PMID: 23630315; PubMed Central PMCID: PMC3695502.
14. Oliveira A, Oliveira LC, Aburjaile F, Benevides L, Tiwari S, Jamal SB, et al. Insight of Genus *Corynebacterium*: Ascertaining the Role of Pathogenic and Non-pathogenic Species. *Frontiers in Microbiology*. 2017;8. doi: ARTN 1937 10.3389/fmicb.2017.01937. PubMed PMID: WOS:000412727900001.
15. Serwold Davis TM, Groman N, Rabin M. Transformation of *Corynebacterium-Diphtheriae*, *Corynebacterium-Ulcerans*, *Corynebacterium-Glutamicum*, and *Escherichia-Coli* with the C-Diphtheriae Plasmid Png2. *Proceedings of the National Academy of Sciences of the United States of America*. 1987;84(14):4964-8. doi: DOI 10.1073/pnas.84.14.4964. PubMed PMID: WOS:A1987J172500067.
16. Tauch A, Kirchner O, Löffler B, Gotker S, Puhler A, Kalinowski J. Efficient electrotransformation of *Corynebacterium diphtheriae* with a mini-replicon derived from the *Corynebacterium glutamicum* plasmid pGA1. *Curr Microbiol*. 2002;45(5):362-7. doi: 10.1007/s00284-002-3728-3. PubMed PMID: WOS:000178144700010.
17. Songer JG, Hilwig RW, Leeming MN, Iandolo JJ, Libby SJ. Transformation of *Corynebacterium-Pseudotuberculosis* by Electroporation. *Am J Vet Res*. 1991;52(8):1258-61. PubMed PMID: WOS:A1991FZ01000011.
18. Garriss G, Poulin-Laprade D, Burrus V. DNA-damaging agents induce the RecA-independent homologous recombination functions of integrating conjugative elements of the SXT/R391 family. *Journal of bacteriology*. 2013;195(9):1991-2003. doi: 10.1128/JB.02090-12. PubMed PMID: 23435980; PubMed Central PMCID: PMC3624592.
19. van Pijkeren JP, Britton RA. High efficiency recombineering in lactic acid bacteria. *Nucleic acids research*. 2012;40(10):e76. doi: 10.1093/nar/gks147. PubMed PMID: 22328729; PubMed Central PMCID: PMC3378904.
20. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*. 2016;44(W1):W16-21. doi: 10.1093/nar/gkw387. PubMed PMID: 27141966; PubMed Central PMCID: PMC4987931.
21. Nakamura Y, Nishio Y, Ikeo K, Gojobori T. The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. *Gene*. 2003;317(1-2):149-55. PubMed PMID: 14604803.
22. Kagawa W, Kurumizaka H, Ishitani R, Fukai S, Nureki O, Shibata T, et al. Crystal structure of the homologous-pairing domain from the human Rad52 recombinase in the undecameric form. *Molecular cell*. 2002;10(2):359-71. PubMed PMID: 12191481.
23. Smith CE, Bell CE. Domain Structure of the Redbeta Single-Strand Annealing Protein: the C-terminal Domain is Required for Fine-Tuning DNA-binding Properties, Interaction with the Exonuclease

Partner, and Recombination in vivo. *Journal of molecular biology*. 2016;428(3):561-78. Epub 2016/01/19. doi: 10.1016/j.jmb.2016.01.008. PubMed PMID: 26780547.

24. Lloyd JA, McGrew DA, Knight KL. Identification of residues important for DNA binding in the full-length human Rad52 protein. *Journal of molecular biology*. 2005;345(2):239-49. Epub 2004/12/02. doi: 10.1016/j.jmb.2004.10.065. PubMed PMID: 15571718.

25. Ploquin M, Bransi A, Paquet ER, Stasiak AZ, Stasiak A, Yu X, et al. Functional and structural basis for a bacteriophage homolog of human RAD52. *Current biology : CB*. 2008;18(15):1142-6. Epub 2008/07/29. doi: 10.1016/j.cub.2008.06.071. PubMed PMID: 18656357; PubMed Central PMCID: PMCPMC4149281.

26. Caldwell BJ, Zakharova E, Filsinger GT, Wannier TM, Hempfling JP, Chun-Der L, et al. Crystal structure of the Redbeta C-terminal domain in complex with lambda Exonuclease reveals an unexpected homology with lambda Orf and an interaction with Escherichia coli single stranded DNA binding protein. *Nucleic acids research*. 2019. Epub 2019/01/10. doi: 10.1093/nar/gky1309. PubMed PMID: 30624736.

27. Lopes A, Amarir-Bouhram J, Faure G, Petit MA, Guerois R. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic acids research*. 2010;38(12):3952-62. doi: 10.1093/nar/gkq096. PubMed PMID: 20194117; PubMed Central PMCID: PMCPMC2896510.

28. Su TY, Jin HY, Zheng Y, Zhao Q, Chang YZ, Wang Q, et al. Improved ssDNA recombineering for rapid and efficient pathway engineering in *Corynebacterium glutamicum*. *J Chem Technol Biot*. 2018;93(12):3535-42. doi: 10.1002/jctb.5726. PubMed PMID: WOS:000451779600019.

29. Sleator RD, Walsh P. An overview of in silico protein function prediction. *Archives of microbiology*. 2010;192(3):151-5. doi: 10.1007/s00203-010-0549-9. PubMed PMID: WOS:000274455600001.

30. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004;32(5):1792-7. doi: 10.1093/nar/gkh340. PubMed PMID: 15034147; PubMed Central PMCID: PMCPMC390337.

31. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research*. 1997;25(24):4876-82. PubMed PMID: 9396791; PubMed Central PMCID: PMCPMC147148.

32. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. *Nucleic acids research*. 2015;43(W1):W389-94. doi: 10.1093/nar/gkv332. PubMed PMID: 25883141; PubMed Central PMCID: PMCPMC4489285.

33. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution*. 2018;35(6):1547-9. doi: 10.1093/molbev/msy096. PubMed PMID: 29722887; PubMed Central PMCID: PMCPMC5967553.

34. Yu X, Jin H, Liu W, Wang Q, Qi Q. Engineering *Corynebacterium glutamicum* to produce 5-aminolevulinic acid from glucose. *Microbial cell factories*. 2015;14:183. doi: 10.1186/s12934-015-0364-8. PubMed PMID: 26577071; PubMed Central PMCID: PMCPMC4650169.

35. Huang Y, Li L, Xie S, Zhao N, Han S, Lin Y, et al. Recombineering using RecET in *Corynebacterium glutamicum* ATCC14067 via a self-excisable cassette. *Scientific reports*. 2017;7(1):7916. doi: 10.1038/s41598-017-08352-9. PubMed PMID: 28801604; PubMed Central PMCID: PMCPMC5554157.

Table 1. The distribution of PRFUs in genus *Corynebacterium*

Species	Genome ^a	SSAP ^a	EXO ^a	PRFU
<i>C. argentoratense</i>	3	1	0	1
<i>C. aurimucosum</i>	6	2	2	1
<i>C. diphtheriae</i>	63	22	22	5
<i>C. falsenii</i>	2	1	1	1
<i>C. freneyi</i>	1	1	1	1
<i>C. glutamicum</i>	20	1	0	1
<i>C. halotolerans</i>	3	1	1	1
<i>C. humireducens</i>	2	2	2	1
<i>C. jeikeium</i>	21	5	5	5
<i>C. kroppenstedtii</i>	4	4	4	3
<i>C. lactis</i>	1	1	1	1
<i>C. lubricantis</i>	1	1	1	1
<i>C. minutissimum</i>	3	1	1	1
<i>C. pyruviciproducens</i>	1	1	1	1
<i>C. simulans</i>	3	1	1	1
<i>C. striatum</i>	10	6	6	4
<i>C. timonense</i>	2	2	2	2
<i>C. ulcerans</i>	17	4	4	4
<i>C. ulceribovis</i>	1	1	1	1
<i>C. variabile</i>	3	1	1	1
<i>C. bouchedurhonense</i>	1	1	1	1
<i>C. phoceense</i>	1	1	1	1
<i>C. provencense</i>	2	1	1	1
<i>C. sp^b</i>	115	22	21	22

^a The number of SSAP or EXO identified or the number of genomes. Species without PRFU

are not listed.

^b Genomes in the genus *Corynebacterium* that haven't been classified.

Table 2. Recombineering with ssDNA using different SSAPs

Strain	Substrate ^a			Viable cells (x10 ⁷) ^a	Transformation efficiency	Recombination efficiency	SSAP type	SSAP source
	ssDNA-	ssDNA+	plasmid (x10 ⁴)					
Sap0	1±1	0	0.67±0.25	41.90±13.11	1.36x10 ⁻⁵	0	-	Empty vector
Sap1	2±3	444±40	2.13±1.16	9.93±3.93	2.15x10 ⁻⁴	2.08x10 ⁻²	SSAP-5	<i>Corynebacterium aurimucosum</i> DSM 44827
Sap2	0	61±30	3.77±1.93	4.03±1.20	9.34x10 ⁻⁴	1.62x10 ⁻³	SSAP-4	<i>Corynebacterium diphtheriae</i> HC07
Sap3	0	136±55	4.21±0.67	16.13±6.38	2.61x10 ⁻⁴	3.23x10 ⁻³	SSAP-3	<i>Corynebacterium glutamicum</i> R
Sap4	0	14±3	2.43±0.41	30.63±7.03	7.92x10 ⁻⁵	5.76x10 ⁻⁴	SSAP-3	<i>Corynebacterium halotolerans</i> 307 CHAL
Sap5	0	0	1.20±0.71	18.10±9.92	6.65x10 ⁻⁵	0	SSAP-1	<i>Corynebacterium kroppenstedtii</i> DNF00591
Sap6	2±2	0	3.26±1.23	41.73±8.27	7.82x10 ⁻⁵	0	SSAP-2	<i>Corynebacterium lubricantis</i> DSM 45231
Sap7	0	0	3.26±0.60	34.00±9.64	9.60x10 ⁻⁵	0	SSAP-2	<i>Corynebacterium striatum</i> 963 CAUR
Sap8	1±2	23±15	2.70±0.43	16.50±2.83	1.64x10 ⁻⁴	8.52x10 ⁻⁴	SSAP-2	<i>Corynebacterium variabile</i> DSM 44702
SmrecT	1±1	20±16	1.19±0.30	22.07±0.98	5.40x10 ⁻⁵	1.68x10 ⁻³	SSAP-5	<i>Escherichia coli</i> MG1655
Sap10	1±1	133±47	3.60±0.85	0.16±0.04	2.30x10 ⁻²	3.69x10 ⁻³	SSAP-3	<i>Corynebacterium sp</i> HMSC070E08
Sap11	-	-	-	-	-	-	SSAP-5	<i>Nocardia farcinica</i> IFM 10152
Sap12	0±0	5±3	8.53±0.92	10.67±1.67	8.00x10 ⁻⁴	5.86x10 ⁻⁵	SSAP-5	<i>Corynebacterium humireducens</i> DSM 45392
Sap13	4±7	16±5	9.33±1.15	10.20±1.40	9.15x10 ⁻⁴	1.71x10 ⁻⁴	SSAP-5	<i>Corynebacterium timonense</i> 5401744
Sap14	0	0	8.40±0.40	9.50±3.25	8.84x10 ⁻⁴	0	SSAP-5	<i>Proteus mirabilis</i> HI4320
Sap15	0	0	9.53±0.50	9.10±1.74	1.05x10 ⁻³	0	SSAP-5	<i>Aneurinibacillus aneurinilyticus</i> ATCC 12856
Sap16	0	0	6.67±1.15	8.50±1.56	7.84x10 ⁻⁴	0	SSAP-5	<i>Thermosipho melanesiensis</i> BI429

^a The cell numbers are shown as the average ± S.E.M. Values indicated are the average of three or more experiments. Transformation efficiency was calculated as the number of cells transformed with plasmid of the viable cells. Recombination efficiency was calculated as the recombinants of the successful transformed cells (cells transformed with plasmid). Sap0 expressing empty vector was used as the negative

449 control. SmrecT expressing RecT was used as positive control. Sap1-Sap8 were tested to verify the function of SSAPs. Sap10-Sap16 were
450 additionally tested to study the relation of SSAP function and local genome context. ssDNA-: no ssDNA control. ssDNA+: adding 2µg oligo
451 substrate. plasmid: adding 500ng plasmid substrate. -: not available.

452

453

454

455

456

457

458

459

460

461

Table 3. Recombineering with dsDNA using different PRFUs

Strain	Substrate ^a			Viable cells (x10 ⁵) ^a	Transformation efficiency	Recombination frequency	PRFU type	PRFU source
	dsDNA-	dsDNA+	plasmid (x10 ⁵)					
ET1	0	2365±822	1.52±0.09	3.73±1.44	4.08x10 ⁻¹	1.56x10 ⁻²	VII	<i>Corynebacterium aurimucosum</i> DSM 44827
ET2	0	4±2	3.11±2.07	6.10±1.87	5.1x10 ⁻¹	1.29x10 ⁻⁵	VI	<i>Corynebacterium diphtheriae</i> HC07
ET3	0	0	1.53±0.50	8.5±1.97	1.8x10 ⁻¹	0	VIII	<i>Corynebacterium glutamicum</i> R
ET4	0	85 ±73	0.66±0.23	11.43±8.44	5.73x10 ⁻¹	1.3x10 ⁻³	IV	<i>Corynebacterium halotolerans</i> 307 CHAL
ET8	-	-	-	-	-	-	III	<i>Corynebacterium variabile</i> DSM 44702
ET9	1	2456 ±391	1.02±0.17	6.03±1.42	1.69x10 ⁻¹	2.41x10 ⁻²	-	<i>Escherichia coli</i> MG1655

^a The cell numbers are shown as the average ± S.E.M. Values indicated are the average of three or more experiments. Transformation efficiency was calculated as the number of cells transformed with plasmid of the viable cells. Recombination efficiency was calculated as the recombinants of the successful transformed cells (cells transformed with plasmid). ET9 expressing RecET was used as the positive control. dsDNA-: no dsDNA control. dsDNA+: adding 1µg dsDNA substrate. plasmid: adding 500ng plasmid substrate. -: not available.

Fig 1. The workflow of finding phage recombinase function unit (PRFU). The genome protein sequence as well as the conserved domain database sequence were first downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). The genome protein sequence database were then searched in the conserved domain database using PSI-BLAST method. Candidate SSAPs or EXOs were then aligned in the conserved domain database to get the consensus parts and form the new conserved domain database. The searching and alignment step were iterated until no new candidates were found. All the candidate SSAPs and EXOs were then collected and paired as PRFUs.

Fig 2. Classification of the PRFUs. **a** Phylogenetic tree and classification of the SSAPs identified. **b** Phylogenetic tree and classification of the EXOs identified. Each type is shown in a different color. The labels are shown as the accession number of the protein, the note for the protein and the source of the protein. The phylogenetic trees were constructed by MEGA6 using Neighbor-Joining method. **c** The schematic second structure of each SSAP types. **d** The N-terminal structure of human SSAP Rad52. PDB :

481 1KN0. **e** The schematic second structure of each EXO types. **f**. The structure of λ Exo. PDB: 6M9K. **g**. The theoretical PRFU type and
482 observed PRFU type. -: not detected.

Fig 3. Recombineering using SSAP with oligo. **a** The plasmid used for SSAP expression. **b** Schematic of oligo mediated recombineering. **c** The culture of no ssDNA control (ssDNA-), no SSAP control (SSAP+) and recombinants survival the selection (SSAP+ ssDNA+). **d** Sequence evidence for four recombinants sample and one no recombineering control.

Fig 5. Local genome context of the PRFUs. **a** The phylogenetic tree of local genome context of PRFU. **b** The local genome context distance (S_{dis}) of different species with PRFUs to that of *C. glutamicum*. PRFUs functioned in the recombineering experiment were colored in green. Those showed no function were colored in red. Additional verified functional SSAPs were colored in orange.

504 **Supporting information**

505 **S1 Fig. The PRFUs in genus *Corynebacterium*.** **a** The genomes with or without PRFUs. **b** The genomes
506 in species with or without PRFU. **c** The phage origin of predicted PRFUs. F: genomes with PRFU, NF:
507 genomes without PRFU. CL: PRFUs found in genomes that has been classified into species, UCL: PRFUs
508 found in genomes that hasn't been classified into any species, F: genomes with PRFU, NF: genomes
509 without PRFU. Numbers indicate the genome number.

510 **S2 Fig. The distribution of different PRFU type among species.** Pathogenetic species were colored in
511 red.

512 **S3 Fig. The 16S rRNA phylogenetic tree of genus *Corynebacterium*.** The 16S rRNA of the representative
513 genomes of classified species were used to generate the tree. Sequences from *Mycobacteria* were
514 used as outgroup. The phylogenetic trees were constructed by MEGA6 (2) using Neighbor-Joining
515 method with the default settings except for the Interior-branch Test value, which is set as 1000.

516 **S4 Fig. Colinear analysis of the local genome context distance and species 16S rRNA distance.** PRFUs
517 functioned in the recombineering experiment were colored in green. Those showed no function were
518 colored in red.

519 **S1 Table. The occurrence of PRFU and genome assemble level**

520 **S2 Table. Strains and plasmids used in the study**