


High-throughput retrieval of target sequences from complex clone libraries using CRISPRi

Received: 27 May 2022

Accepted: 28 September 2022

Published online: 21 November 2022

 Check for updates

Ján Burian, Vincent K. Libis , Yozen A. Hernandez, Liliana Guerrero-Porras, Melinda A. Ternei  & Sean F. Brady  

The capture of metagenomic DNA in large clone libraries provides the opportunity to study microbial diversity that is inaccessible using culture-dependent methods. In this study, we harnessed nuclease-deficient Cas9 to establish a CRISPR counter-selection interruption circuit (CCIC) that can be used to retrieve target clones from complex libraries. Combining modern sequencing methods with CCIC cloning allows for rapid physical access to the genetic diversity present in natural ecosystems.

CRISPR–Cas systems can be programmed to target essentially any unique DNA or RNA sequence when loaded with a homologous guide RNA¹, and they have been adapted for numerous genetic engineering and synthetic biology tools^{2,3}. Cas9 is a dual-RNA-guided DNA nuclease that binds a trans-activating crRNA (tracrRNA) base paired to a target-encoding CRISPR RNA (crRNA) forming an active complex that can target 20-bp DNA regions that contain a 3' NGG PAM site⁴. CRISPR interference (CRISPRi) exploits the sequence-specific localization of a nuclease-deficient Cas (for example, dCas9) to repress gene expression by blocking either access to a promoter or transcript elongation.⁵ Recently, Cas9 and dCas9 have been adapted as enrichment tools for targeted sequencing⁶, enrichment of mutants within a heterogeneous population⁷ and genotypic enrichment during chemical-genetic profiling^{8,9}. Although dCas9 has generally been targeted to either endogenous genomic loci or a limited set of genetic circuits, we reasoned that incorporating a degenerate target sequence (barcode) into a synthetic circuit would generate a pool of unique constructs that could be selectively targeted by dCas9 to trigger clone-specific tasks. By placing a counter-selection marker under the control of such a circuit, silencing a specific barcode sequence would lead to target retrieval (that is, survival) under selective conditions. The development of a method for rapid and high-throughput retrieval of specific sequences from large clone libraries would accelerate the discovery of novel genes and biosynthetic gene clusters (BGCs). This is particularly true when exploring microbial communities using culture-independent (for example, metagenomic) methods where metagenomic libraries can contain tens of millions of unique clones due to the immense diversity present in natural ecosystems. The targeted retrieval of clones from large metagenomic libraries has been limited to a laborious

multi-step dilution and polymerase chain reaction (PCR) screening method¹⁰, which remains a key bottleneck in discovery pipelines/platforms. Although CRISPR–Cas has been used to streamline a number of methods for the precise cloning of target genomic sequences^{11–13}, these methods have not been adapted to access sequences from complex metagenomic libraries. In this study, we developed a barcoded CRISPR counter-selection interruption circuit (CCIC) that we combined with two advanced sequencing methods—PacBio long-read sequencing¹⁴ and edge mapping¹⁵—for rapid indexing and retrieval of target sequences from complex metagenomic and genomic libraries.

Targeting dCas9 to a sequence between a counter-selection gene and its promoter should abrogate expression and allow for survival under otherwise selective conditions (Fig. 1a). To test the feasibility of a CCIC, we used a pair of P1-derived artificial chromosome (PAC)¹⁶ vectors that were identical except for their multiple cloning sites (MCSs) located between *sacB* and its strong constitutive promoter (Fig. 1b). This allowed us to design a guide RNA homologous to a sequence within the pPAC-T MCS (that is, target) that was not present in the MCS of pPAC-N (that is, negative). *SacB* causes cell death by producing toxic levan in the presence of sucrose¹⁷. Sequence-specific survival on sucrose was observed only when a plasmid expressing dCas9, tracrRNA and target-specific crRNA was provided (Fig. 1c). The highest level of survival was observed using *Escherichia coli* with a genomically integrated *dcas9* transformed with a guide plasmid expressing a tracrRNA/crRNA chimera (single guide RNA (sgRNA); Fig. 1c). We tested whether this system could be used to recover a single target PAC present in populations of 5,000, 50,000 and 100,000 non-target PACs. Efficient retrieval of the target PAC (positive hit rate >70%) was achieved in mixtures of up to 50,000 non-target sequences (Fig. 1d).

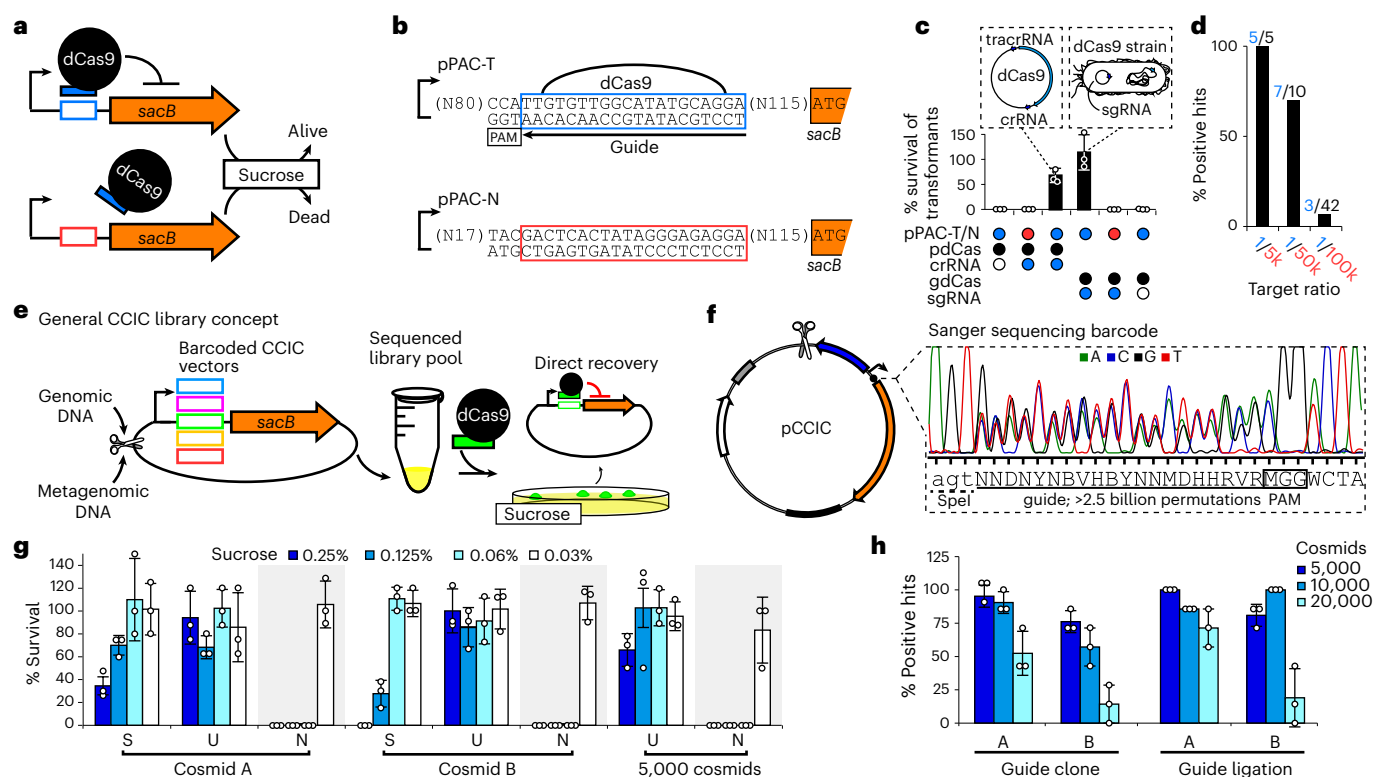


Fig. 1 CCIC development and its application for target sequence retrieval.

a, CCIC concept: dCas9 loaded with a guide RNA (solid blue box) targeting a sequence downstream of *sacB*'s promoter (open blue box) will lead to repression, allowing for survival on sucrose media. A non-target sequence is not recognized (open red box), leading to cell death upon sucrose exposure. **b**, A guide RNA corresponding to pPAC-T (blue), but not in pPAC-N (red), was designed to test the feasibility of selective dCas-mediated *sacB* repression. **c**, Effectiveness of dCas9-mediated *sacB* repression measured as % survival of transformants containing pPAC-T (blue) or pPAC-N (red) on sucrose-containing media relative to non-sucrose media. dCas9 components were provided on a plasmid (pdCas9), or sgRNA was expressed in a strain with genomically incorporated dCas9 under its native promoter (gdCas; *E. coli* NdC). **d**, Direct retrieval of pPAC-T when mixed with pPAC-N at different ratios. Number of pPAC-T colonies (blue) are identified, and the pPAC-N colonies screened (black) are indicated. **e**, General CCIC library concept. Genomic or metagenomic DNA is cloned into a collection of barcoded

CCIC vectors to generate a (meta)genomic library. The resulting library is sequenced to link a unique barcode with each DNA insert. A target clone can then be retrieved by dCas9-mediated repression of *sacB* (orange) using barcode (open green box)-specific guide RNAs (solid green box) and plating on sucrose media. **f**, Plasmid map of pCCIC containing a degenerate barcode validated by Sanger sequencing. **g**, Silencing of barcoded metagenomic cosmids (cosmid 'A', cosmid 'B' and pool of '5,000' cosmids) at various sucrose concentrations using barcode-specific sgRNAs (S), universal NP1 sgRNA (U) or a non-specific sgRNA (N). *E. coli* JdC, containing genomically integrated dCas9 with a strong constitutive promoter, was used. **h**, Direct retrieval of cosmid 'A' and 'B' from cosmid libraries of different sizes using 0.07% sucrose counter-selection. Isolated guide plasmids or guide plasmid ligations specific for cosmid 'A' and 'B' barcodes were tested for retrieval. Data in **c**, **g** and **h** are presented as mean values \pm standard deviation ($n = 3$ independent transformations and screenings per retrieval).

The target PAC was also retrieved from the 100,000 non-target mixture, albeit with ten-fold-reduced efficiency (Fig. 1d). These studies, using a model two-vector mixture, confirmed the potential for a CCIC to allow targeted retrieval of sequences from complex mixtures.

Cosmid library construction using lambda phage packaging offers a simple method for large-scale capture of high-molecular-weight metagenomic DNA fragments¹⁸. To test CCIC-based sequence retrieval from large-insert clone libraries, we developed a CCIC-containing cosmid vector. Among the various *sacB* promoter and replication origin options we tested, the combination of TetR-repressed *sacB* carried on a pMB1 backbone showed the best overall CCIC performance (Supplementary Fig. 1). These components were, therefore, used to construct the cosmid cloning vector pCCIC. Lambda phage efficiently packages DNA fragments from 37 kb to 52 kb in size¹⁹, and so libraries constructed with the 6-kb pCCIC are expected to contain 31–46-kb metagenomic inserts (Supplementary Table 1; average insert of 35.6 ± 0.4 kbp). CCIC-based retrieval relies on each vector within a library containing a unique sequence between *sacB* and its promoter. This sequence acts both as a barcode for the captured DNA and a guide RNA target for dCas9-mediated repression (Fig. 1e). This was achieved by introducing a degenerate 24-bp barcode/dCas9-targetable sequence into pCCIC

(Fig. 1f). Generating a large pool of unique barcode sequences located between *sacB* and its promoter was achieved by introducing a degenerate 24-bp dCas9-targetable sequence into pCCIC using two-fragment cloning (Fig. 1f and Supplementary Fig. 2a). Sequencing of a (meta)genomic library constructed using a collection of barcoded pCCIC vectors will link specific captured sequences with the unique vector barcodes. Any clone within the library can then be easily retrieved by barcode-specific dCas9-mediated inhibition of *sacB* expression, leading to clone-specific survival on sucrose-containing media (Fig. 1e). We found that addition of the barcode invariably led to sucrose escape mutants, with the highest fidelity cloning (two-fragment ligation) leading to ~0.15% escapes (Supplementary Fig. 2a). Sucrose escape mutants will result in false-positive colonies appearing on selective media, thereby increasing the number of colonies that must be screened to identify a desired target during retrieval. As an example, at 0.15% one would expect 15 false positives when retrieving a target from a 10,000-member library. To generate high-fidelity pools of CCICs, we developed a procedure where sub-pools of clones were grown, checked for CCIC integrity and then pooled to generate 'scrubbed' pools with substantially reduced escape frequencies (Supplementary Fig. 2b; fidelity >99.999%). With the fidelity improvement, one would now

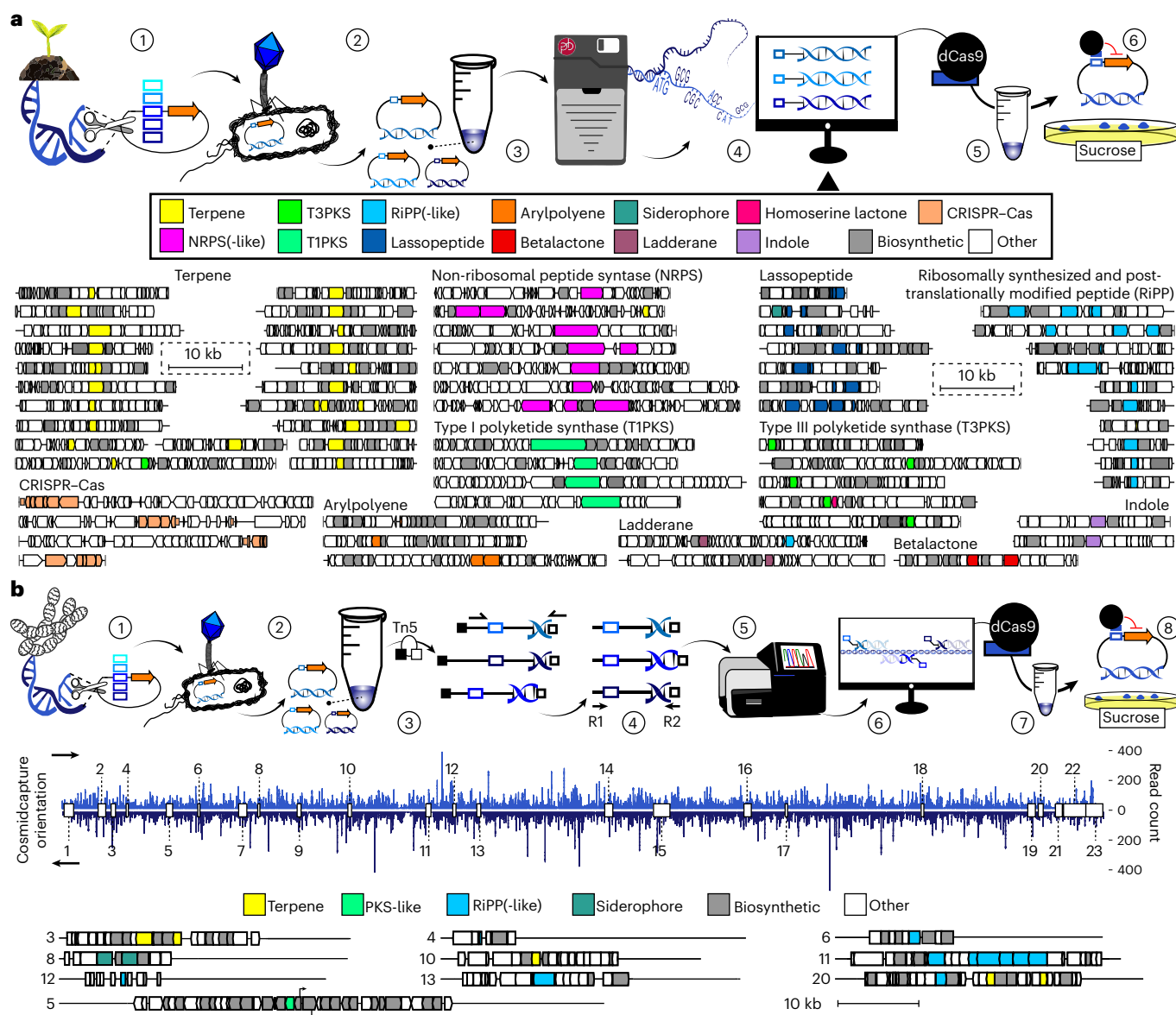


Fig. 2 | Targeted retrieval from metagenomic and genomic CCIC libraries.

a, Top: general outline of metagenomic mining using CCIC retrieval. (1) Soil-extracted DNA is cloned into barcoded pCCIC (barcodes are blue gradient boxes; *sacB* is in orange) using lambda phage packaging (2) to create a cosmid library in *E. coli*. (3) The cosmid library is sequenced by PacBio HiFi CSS. (4) Bioinformatic analysis of the assembled sequence data identifies captured diversity and vector barcodes. (5) Guide RNA matching a desired barcode is transformed into the library pool (6) triggering target-specific dCas9 silencing of *sacB*, leading to target retrieval by clone-specific survival on sucrose. Bottom: CCIC retrieval was used to isolate four CRISPR–Cas systems and a diverse collection of BGCs representing 12 different major biosynthetic classes. **b**, Top: general pipeline for genomic library mining using edge mapping and CCIC. (1) Genomic DNA is ligated into barcoded pCCIC vectors (barcodes are blue gradient boxes; *sacB* is in orange) by lambda phage packaging (2) to generate a cosmid library. (3) Library DNA is fragmented using Nextera ‘tagmentation’ (that is, Tn5 transposase),

allowing for PCR amplification of fragments containing both a vector barcode and the edge of the cloned sequence. (4) Sequencing-ready amplicons are generated allowing for (5) paired-end MiSeq reads to link barcodes and the edge sequences. (6) As lambda phage captures 30–40 kb of sequence, these data generate a comprehensive index of captured regions across a reference genome. (7) A guide RNA matching a desired barcode linked to a target genomic region is transformed into the library pool triggering target-specific dCas9 silencing of *sacB* and (8) leading to target retrieval by clone-specific growth on sucrose. Middle: edge mapping data from an 11,000-member *S. albidoflavus* cosmid library overlaid on the reference genome annotated with the location of 23 BGCs. Bottom: All previously uncharacterized BGCs that could fit on a single cosmid were isolated using edge mapping and CCIC retrieval. The precision of edge mapping also allowed us to isolate two overlapping cosmids that contained a 41-kb polyketide synthase BGC (#5). Arrows indicate the edge of each cosmid.

expect 0.1 false-positive colonies for every on-target colony found using CCIC retrieval from a 10,000-member library. dCas9-mediated silencing of *sacB* expression by targeting pCCIC barcodes was then optimized by increasing dCas9 expression levels and screening the sucrose concentration used for selection (Supplementary Fig. 3).

To investigate CCIC utility for isolating target sequences from complex mixtures, we generated a cosmid-based library from soil

metagenomic DNA and attempted to recover two randomly selected clones. Barcodes from these clones were sequenced and used to generate guide plasmids expressing homologous sgRNAs. A range of sucrose concentrations was tested for optimal dCas9-mediated survival of the individual barcoded clones using their specific guide plasmids (Fig. 1g: S). Fidelity of the CCIC in individual clones, as well as a pool of 5,000 metagenomic clones, was confirmed by silencing with a universal

guide plasmid that targets a constant sequence in the vector (Fig. 1g:U) and by lack of survival on sucrose with non-specific guides (Fig. 1g:N). Using our final CCIC selection protocol, the randomly chosen barcoded clones were easily retrieved from pools of up to 20,000 metagenomic clones using their corresponding guide plasmids (Fig. 1h). In these initial studies, we used purified guide plasmids for each retrieval experiment. In an effort to further simplify the CCIC method, we tested whether the ligation reactions that generated the guide plasmids could be used directly for retrieval. As seen when using purified guide plasmid clones, direct transformation of the ligation reactions resulted in successful target retrieval from pools containing as many as 20,000 distinct clones (Fig. 1h). These studies indicated that barcoded CCIC clone libraries could provide a simple and rapid way to recover targeted DNA sequence from a complex mixture.

We next used the CCIC method to recover potentially high-value sequences of interest (that is, natural product BGCs and CRISPR–Cas systems) from the 10,000-clone metagenomic library used in Fig. 1h. To link barcodes with specific cosmid inserts, we sequenced the library pool using PacBio HiFi long-read technology.¹⁴ Assembled contigs containing barcodes were analyzed for BGC content and phage-defense systems using the publicly available prediction tools antiSMASH²⁰ and DefenseFinder²¹, respectively. Based on these analyses, 66 cosmids predicted to contain 12 different classes of core biosynthetic genes and four cosmids containing CRISPR–Cas genes were selected for retrieval. Barcodes from the 70 target cosmids were used to design sgRNAs, and direct transformation of the guide plasmid ligation (like Fig. 1h) allowed us to recover over 95% of our desired targets (Fig. 2a). Guide RNAs are known to have varied activities²², with low-efficacy guides possibly contributing to the failure of some retrieval attempts. A ‘postmortem analysis’ of the CCIC method indicated that the positive hit rate for a retrieval likely depended on a combination of guide strength, cosmid copy number and varied clone abundance within the library (Supplementary Fig. 4 and Supplementary Note 1). Multiple mechanisms of sucrose escape were observed upon sequencing a collection of false-positive clones (Supplementary Table 2 and Supplementary Note 1). All desired targets could likely be retrieved from a library with increased coverage (that is, as few as two available barcodes for each target). From receiving primers to recovery of the clone of interest, CCIC retrieval requires just 2 days, which is a marked increase in efficiency compared to all other methods that we have explored (Supplementary Note 2).

As an alternative to linking inserts to barcodes using long-read sequencing, the proximity of the barcode to one edge of the DNA captured in pCCIC led us to develop a PCR-based ‘edge mapping’ method that can be used to index a complex clone library with minimal effort and sequencing resources. Although several methods exist to precisely extract sequences from sequenced genomes²³, the large-scale parallel cloning of target genomic regions remains cumbersome. As genomic libraries are created in a sequence-independent manner, they can easily capture all of the encoded diversity with only the bottleneck of a laborious screening step limiting the rate of target sequence retrieval²³. For edge mapping, Tn5 transposase tagmentation²⁴ was used to insert known sequence tags upon DNA fragmentation, which allowed us to amplify fragments containing the vector barcode and the edge of each cloned sequence by PCR. Paired-end MiSeq reads then linked the barcode to the edge sequence, generating a comprehensive index of captured regions with the assumption that lambda phage packaging captured the expected 31–46 kb of sequence (Fig. 2b). Paired with CCIC, this allowed for high-throughput retrieval of specific genomic loci.

To demonstrate the utility of edge mapping for BGC retrieval, we generated a high-density cosmid library (~11,000 clones = ~195× genome coverage) from the genomic DNA of *Streptomyces albidoflavus* J1074, a representative BGC-rich Actinomycete. Edge sequences from this library were linked to a total of 10,145 unique barcodes, mapped to the *S. albidoflavus* genome, and clones predicted to contain nine

uncharacterized BGCs that could each be carried on a single cosmid were identified and retrieved from the library by CCIC (Fig. 2b). With the exception of one BGC, all target BGCs were retrieved on the first attempt. Due to the saturation level of the library (Supplementary Fig. 5), a second barcode/guide was identified and used to successfully recover the final BGC. As the edge mapping produced base pair resolution of captured edges, it was possible to identify precise sets of overlapping cosmids for BGCs too large to capture on a single cosmid. As an example, we used our mapping data to recover two overlapping cosmids that contained a polyketide synthase BGC (Fig. 2b). Overall, these data demonstrated the utility of CCIC to easily scale the capture and retrieval of target sequences from sequenced genomes (Supplementary Note 3). In the case of metagenomic libraries, edge mapping should be particularly useful for indexing large libraries to guide the identification of overlapping clones (Supplementary Fig. 6a), which has been a key bottleneck in the cloning of large complete metagenomic BGCs that require multiple cosmids to fully assemble. Alignment of edge mapping data from our 10,000-member metagenomic library to the PacBio assemblies predicted several instances of potentially overlapping cosmids. We confirmed one of the predictions by retrieval of cosmids associated with PacBio contig 1,912, including the edge barcode (barcode operational taxonomic unit (OTU) 477) and the internally mapped barcode OTU 315 (Supplementary Fig. 6b and Supplementary Note 2). We also found that edge mapping provides a cost-effective means of correcting low-quality barcode sequences found in long-read sequencing datasets (Supplementary Fig. 7 and Supplementary Note 2).

As next-generation sequencing methods have increasingly provided unprecedented bioinformatic access to genetic diversity²⁵, methods for physically accessing sequences of interest have remained rudimentary. By harnessing the vast target potential of dCas9, CCIC cloning opens the door to rapid, scalable and cost-effective indexing and retrieval of target sequences. Although we applied CCIC to accelerate (meta)genomic mining, we are excited to see how the general concept of incorporating degenerate Cas-targetable barcodes into genetic circuits is expanded into other areas of synthetic biology in the future.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-022-01531-8>.

References

1. Wang, J.Y., Pausch, P. & Doudna, J.A. Structural biology of CRISPR–Cas immunity and genome editing enzymes. *Nat. Rev. Microbiol.* **20**, 641–656 (2022).
2. Xu, X. & Qi, L. S. A CRISPR–dCas toolbox for genetic engineering and synthetic biology. *J. Mol. Biol.* **431**, 34–47 (2019).
3. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).
4. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
5. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR–Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).
6. Schultzhau, Z., Wang, Z. & Stenger, D. CRISPR-based enrichment strategies for targeted sequencing. *Biotechnol. Adv.* **46**, 107672 (2021).
7. Feldman, D. et al. CloneSifter: enrichment of rare clones from heterogeneous cell populations. *BMC Biol.* **18**, 177 (2020).

8. Li, S. et al. CRISPRi chemical genetics and comparative genomics identify genes mediating drug potency in *Mycobacterium tuberculosis*. *Nat. Microbiol.* **7**, 766–779 (2022).
 9. Jost, M. et al. Combined CRISPRi/a-based chemical genetic screens reveal that rigosertib is a microtubule-destabilizing agent. *Mol. Cell* **68**, 210–223 (2017).
 10. Owen, J. G. et al. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. *Proc. Natl Acad. Sci. USA* **112**, 4221–4226 (2015).
 11. Jiang, W. et al. Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat. Commun.* **6**, 8101 (2015).
 12. Lee, N. C., Larionov, V. & Kouprina, N. Highly efficient CRISPR/Cas9-mediated TAR cloning of genes and chromosomal loci from complex genomes in yeast. *Nucleic Acids Res.* **43**, e55 (2015).
 13. Wang, H. et al. ExoCET: exonuclease in vitro assembly combined with RecET recombination for highly efficient direct DNA cloning from complex genomes. *Nucleic Acids Res.* **46**, e28 (2018).
 14. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
 15. Burian, J. & Thompson, C. J. Regulatory genes coordinating antibiotic-induced changes in promoter activity and early transcriptional termination of the mycobacterial intrinsic resistance gene *whiB7*. *Mol. Microbiol.* **107**, 402–415 (2018).
 16. Pierce, J. C., Sauer, B. & Sternberg, N. A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: improved cloning efficacy. *Proc. Natl Acad. Sci. USA* **89**, 2056–2060 (1992).
 17. Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T. & Kado, C. I. Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. *J. Bacteriol.* **164**, 918–921 (1985).
 18. Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* **2**, 1297–1305 (2007).
 19. Haley, J. D. in *New Nucleic Acid Techniques* (ed Walker, J. M.) 257–283 (Humana Press, 1988).
 20. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
 21. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
 22. Calvo-Villamanan, A. et al. On-target activity predictions enable improved CRISPR–dCas9 screens in bacteria. *Nucleic Acids Res.* **48**, e64 (2020).
 23. Wang, W., Zheng, G. & Lu, Y. Recent advances in strategies for the cloning of natural product biosynthetic gene clusters. *Front. Bioeng. Biotechnol.* **9**, 692797 (2021).
 24. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
 25. Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C. & Scorilas, A. Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. *Life (Basel)* **12**, 30 (2021).
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
- © The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Methods

Strain and vector construction

Construction of strains and vectors used in this study are detailed in the Supplementary Methods. A strain table including previously published strains and constructs^{26–32} is provided as Supplementary Table 3, and the primers used for constructions in Supplementary Table 4.

Barcoding

The CCIC system depends on a unique dCas9-targetable sequence being present between *sacB* and its promoter. Based on Cas9 preference data from Doench et al.³³, which focused on the rational design of highly active sgRNA for Cas9 cleavage, we designed the degenerate sequence NNDNNNNHHNNHHHHVVRvgh that we thought could serve this role and minimize poorly targeted sgRNA sequences. We then modified this sequence to eliminate the presence of SmaI, SpeI and MscI restriction sites, resulting in the final degenerate sequence NNDNYNBVHBYNNMDHHRVRmggw. This sequence served as the barcode/sgRNA target. Two methods were tested for barcode addition: whole-vector PCR followed by SpeI digest and self-ligation and two-fragment ligation. The two-fragment ligation showed the lowest CCIC escape frequency (Supplementary Fig. 2) and was, therefore, used for the large-scale barcode cloning method described below.

Barcode sequences were added by restriction cloning using SpeI and MscI, where the primer BCode_Msc_R was paired with degenerate sequence-appended primers to generate a barcoded amplicon. The primers BCode_SpeI_F and BCode_NP1 were used to barcode pWEB. tS-MA for experiments presented in Supplementary Fig. 3b and Supplementary Fig. 3d,e, respectively, whereas the primer BCode_Final was used to barcode pCCIC for library construction. In each case, the non-barcode vector was used as the template for PCR. For large-scale barcoding, multiple PCR reactions were pooled and purified. The purified PCR product was polished with T4 DNA polymerase (New England Biolabs (NEB)) per the manufacturer's instructions and once again column purified. The column eluent was digested with SpeI and MscI and then gel purified. Each vector was digested with SpeI and MscI, and the backbone fragment was gel purified. The final polished, digested and gel-purified barcoded PCR product was ligated with the purified vector backbone fragment in 20 μ l of ligation reactions with T4 DNA ligase (NEB) (16 hours, -22°C). Each reaction contained 120 ng of the barcoded amplicon and 50 ng of vector backbone. Ligations were transformed into TransFormMax EC100 (Lucigen EC10010) electrocompetent cells and recovered in LB shaking (200 r.p.m.) at 37°C for 45 minutes. Aliquots of transformations were plated on LB with chloramphenicol to estimate the cloned titer, and the remaining culture was added to LB with chloramphenicol at a 1:40 ratio to amplify the barcoded vector library. Ligations were scaled as necessary to produce the desired quantity of barcoded vector.

Barcode scrubbing

The barcoded pool of pCCIC vectors contained clones that escaped sucrose counter-selection. Although the escape frequency was low ($-1/660$), it was enough to interfere with the efficiency of targeted recovery from large clone pools. To decrease the rate of sucrose escape mutants, we developed a procedure where sub-pools of clones were grown, checked for CCIC integrity and then pooled to generate 'scrubbed' barcoded pCCIC vector pools. A barcoded vector pool or cosmid library was grown overnight to confluence in LB + chloramphenicol at 37°C and 200-r.p.m. shaking. The optical density at 600 nm (OD_{600}) was measured, and the culture was diluted to a titer of ~ 200 cells per 50 μ l in fresh LB + chloramphenicol. This dilution was based on the assumption that an OD_{600} of 1 corresponded to 3×10^8 colony-forming units per milliliter (CFU ml^{-1}) of *E. coli*. Then, 25 ml of diluted culture was prepared for each 384-well microplate to be seeded. A 12-channel pipette was used to seed each well in a 384-well microplate (VWR, 781281) with 50 μ l of the diluted cells. The

plate(s) were then grown overnight at 37°C and 400-r.p.m. shaking. To confirm the titer of the dilute cell preparation, three LB + chloramphenicol agar plates were spread with 50 μ l of the dilution and grown overnight at 37°C . Colonies were counted the next day to generate an exact titer for each experiment. LB agar with chloramphenicol, 1% sucrose and 100 ng ml^{-1} of anhydrotetracycline was prepared in OmniTrays (Thermo Fisher Scientific, 242811) that match the shape of the 384-well microplates. A 384-pin multi-blot replicator (VP 384, V&P Scientific) that delivers ~ 0.2 μ l was used to replica-plate the microplate cultures onto the agar OmniTrays. The pinned OmniTrays were then grown overnight at 37°C , and the microplates were stored at 4°C . After overnight incubation, the OmniTrays were examined to determine which microplate wells did not generate any visible growth. The 'no-growth' wells were combined to generate a 'scrubbed' pool. The scrubbing process is illustrated with example results in Supplementary Fig. 2b. Although likely compatible with automation, this scrubbing process was designed to generate robust libraries by hand without requiring expensive specialized tools, thus making it accessible to as many laboratories as possible. In the course of multiple experiments to establish the CCIC method, a single postdoctoral researcher was able to scrub more than 1 million clones using the described method. For the metagenomic library construction, a scrubbed vector pool of $\sim 350,000$ clones was generated. We note that the escape frequency increased after the metagenomic DNA cloning step, implying that each cloning reaction generates CCIC escape constructs. Therefore, for the *Streptomyces* genomic DNA library construction, we used the 20-million-barcode pool for library construction and scrubbed the library after cloning.

Preparation of CaCl_2 competent cells and transformation

The protocols for generating CaCl_2 competent cells and transformation were as described by Chan et al.³⁴ with minor modifications. Competent cell aliquots were incubated with DNA on ice for 30 minutes, followed by heat-shock at 45°C for 45 seconds, ice incubation for 1 minute and recovery in 1 ml of LB, followed by the addition of 1 ml of LB broth. The tube was then placed in a 37°C shaking (200 r.p.m.) incubator for 30-minute recovery. Recovery times for CCIC are described below.

Sucrose survival assays and direct recovery

Spotting assay. To investigate the survival elicited by dCas9 silencing of *sacB* expression (presented in Fig. 1), we spotted a serial dilution of *E. coli* transformations on selective and non-selective media and compared the CFU. After guide RNA was transformed into target competent cells and the subsequent recovery (described below), a ten-fold serial dilution series was prepared from no dilution up to 10^{-7} . Then, 6 μ l of each dilution was spotted onto selective and non-selective media, and the plates were incubated overnight at 37°C . Colony counts were then performed, and the survival percentage (colonies on selective media versus non-selective media) was calculated.

Recovery conditions. Vectors tested in Fig. 1c,d had a recovery of 1-hour shaking (200 r.p.m.) at 37°C , after which spectinomycin and chloramphenicol were added, and the recovery continued shaking (200 r.p.m.) at 37°C for an additional 1 hour. The selective plates contained 5% sucrose and 50 $\mu\text{g L}^{-1}$ of aTc. Derivatives of pCCIC (tested in Fig. 1g,h and Fig. 2) were recovered for 1-hour shaking (200 r.p.m.) at 37°C . Selective plates contained 100 $\mu\text{g L}^{-1}$ of aTc and the indicated sucrose concentrations or, for the mass recoveries in Fig. 2, 0.07% sucrose.

Screening clones. To rapidly screen constructs (pPAC-T or pPAC-N) isolated in the recovery experiments presented in Fig. 1d, colony PCR was carried out using primers pPACMCSeq_F and pPACMCSeq_R. pPAC31-N would generate a 184-bp band, whereas pPAC31-T would generate a 247-bp band. These differences were well-resolved on a 2%

agarose gel. Colonies were suspended in 30 μ l of water, and 1 μ l of the suspension was used as template for a 25- μ l PCR reaction using Q5 polymerase as per the manufacturer's instructions. Sanger sequencing of representative clones, using the sacBout primer, was used to confirm the PCR results. To rapidly screen barcoded vectors, colony PCR was performed using the BC_pullR primer paired with the sgRNA/barcode sequence. For Fig. 1h primers, 350A and 350B were used. Colonies were screened as described above, with the following minor modifications. Due to the large number of clones to be screened, Syto 9 (Thermo Fisher scientific) was added to the PCR mixture (0.0025 μ l per 25 μ l), and post-PCR melt curve analysis (0.5 $^{\circ}$ C increments from 75 $^{\circ}$ C and 95 $^{\circ}$ C) was carried out using a CFX384 Touch Real-Time PCR Detection System paired with a C1000 Touch thermal cycler (Bio-Rad). We confirmed that a melt curve peak at -84 $^{\circ}$ C was diagnostic of a positive hit by Sanger sequencing representative clones with the sacBout primer. The presence of this peak was used, at scale, to determine correctly recovered clones.

CCIC recovery

As described above for the rapid screening of barcoded vectors, CCIC-recovered colonies were screened by pairing the BC_pullR primer with target barcode sequence primer. The primers used for the metagenomic and genomic library recoveries are listed in Supplementary Tables 5 and 6, respectively. To ensure that large-scale screening was as cost-effective as possible, 10- μ l PCR reactions with Taq were used. For each screen, colonies were suspended in 30 μ l of water, and 1 μ l of the suspension was added to a 10- μ l PCR reaction as described below:

Component	μ l
Water	6.7
Bulldog Buffer	1
5M betanine	1.6
10mM dNTP	0.2
50 μ M primer F	0.2
50 μ M primer R	0.2
Taq	0.1
Syto 9	0.005

PCRs were run in 384-well plates using a CFX384 Touch Real-Time PCR Detection System paired with a C1000 Touch thermal cycler (Bio-Rad) and the follow cycling conditions: 94 $^{\circ}$ C for 2 minutes, 34 cycles of 94 $^{\circ}$ C for 30 seconds, 55 $^{\circ}$ C for 30 seconds, 72 $^{\circ}$ C for 30 seconds and then a final extension of 1 minute at 72 $^{\circ}$ C. A melt curve with 0.5 $^{\circ}$ C increments from 75 $^{\circ}$ C and 95 $^{\circ}$ C was performed after PCR. Clones containing the desired barcode had a melt curve peak at -80 $^{\circ}$ C (shifted from -84 $^{\circ}$ C due to the change in the PCR buffer). Sanger sequencing of the barcode using the sacBout primer and the edge of the insert, using the M13-40FOR primer, were used to confirm isolation of the target clone. For large-scale recovery efforts, eight colonies were screened per barcode target. If no correct peak was identified, an additional 24 colonies were screened. In general, most (>70%) targets were identified with screening eight colonies.

Soil metagenomic DNA isolation

Soil metagenomic DNA was extracted according to previously published protocols¹⁸.

S. albidoflavus J1074 genomic DNA extraction

S. albidoflavus J1074 was shaken (200 r.p.m.) in 30 ml of TSB within a baffled 125-ml flask containing a 3-cm³ piece of chicken wire at 30 $^{\circ}$ C until saturation. One milliliter of culture was centrifuged in a 2-ml

Eppendorf tube, washed once with 500 μ l of GTE (25 mM Tris-HCl pH 8, 10 mM EDTA and 50 mM glucose) and re-suspended in 500 μ l of GTE containing 1 mg ml⁻¹ of lysozyme. The suspension was incubated for 60 minutes at 37 $^{\circ}$ C, followed by the addition of 50 μ l of 10 mg ml⁻¹ Proteinase K and 100 μ l of 10% SDS. After mixing by inversion, the sample was incubated at 55 $^{\circ}$ C for 60 minutes. Then, 200 μ l of 5 M NaCl and 160 μ l of Na-CTAB (mixture of 4.1% w/v NaCl and 10% w/v hexadecyltrimethylammonium bromide) were added; the sample was mixed by pipetting and then placed at 65 $^{\circ}$ C for 10 minutes. An equal volume (1 ml) of chloroform-isoamyl alcohol (24:1) was then mixed in by pipetting, and the sample was centrifuged at 24,000 g for 5 minutes. The top aqueous layer (~800 μ l) was transferred to a 1.5-ml Eppendorf tube; 560 μ l of isopropanol was added; and the genomic DNA was left to precipitate at -22 $^{\circ}$ C for 5 minutes. The DNA was then pelleted by centrifugation at 24,000 g for 15 minutes, washed once with 500 μ l of 70% ethanol and air dried for 20–30 minutes. Finally, the extracted genomic DNA was re-suspended in 250 μ l of water with heating at 50–55 $^{\circ}$ C.

Lambda phage packaging

Vectors for packaging were digested with SmaI and dephosphorylated using Quick CIP (NEB) according to the manufacturer's instructions. DNA extracts were blunted using the End-It DNA End-Repair Kit (Lucigen) according to the manufacturer's instructions. A 5- μ l ligation reaction with 125 ng of blunted insert DNA and 250 ng of vector using the Fast-Link DNA Ligation Kit (Lucigen) was prepared and incubated overnight at -22 $^{\circ}$ C. The ligation was then used for lambda phage packaging with MaxPlax Lambda Packaging Extracts (Lucigen) according to the manufacturer's instructions. For the metagenomic library, multiple pools of ~5,000 metagenomic clones each were generated from the 350,000 scrubbed barcoded pCCIC vector pool and mixed together as needed. For the *S. albidoflavus* library, the 20 million barcoded 'unscrubbed' pCCIC vector pool was used to construct an estimated ~35,000-cosmid pool, from which a ~11,000 library was 'scrubbed'.

Long-read sequencing

The pool of 10,000 metagenomic cosmid clones was plasmid prepped, and the extract was treated with plasmid safe (Lucigen) overnight, according to the manufacturer's instructions, to remove sheared cosmids and *E. coli* genomic contamination. The sample was then submitted to the Vertebrate Genomes Laboratory (Rockefeller University) for PacBio library preparation and HiFi sequencing.

Edge mapping

The method was developed based on a previously established Tn-Seq protocol¹⁵. Nextera XT tagmentation was performed as per the manufacturer's instructions (10 μ l of TD, 5 μ l of DNA and 5 μ l of ATM) with a total of 2 ng of the scrubbed library. Three tagmentation reactions were run in parallel. The DNA Clean & Concentrator-5 (Zymo Research) with a 5:1 ratio of DNA binding buffer was used to pool, purify and concentrate the three tagmentation reactions into a 10- μ l elution volume. A 100- μ l master PCR mix was prepared with 36 μ l of water, 50 μ l of Buffer G (Lucigen), 10 μ l of eluate, 2 μ l each of the P7tag and BC_F primers at 20 μ M and 1 μ l of Taq (Bulldog Bio). The master mix was split into five 20- μ l PCR reactions. PCR was performed using the following conditions: 95 $^{\circ}$ C for 1 minute, followed by 18 cycles of 95 $^{\circ}$ C for 30 seconds, 55 $^{\circ}$ C for 30 seconds and 72 $^{\circ}$ C for 90 seconds. The PCR reactions were then separated on a 1.5% agarose gel, and the smear between 1 kb and 1.5 kb was purified. The DNA concentration of the 20- μ l elution was determined using a Qubit dsDNA high-sensitivity assay. Then, 0.6 ng of DNA was used to seed 20 μ l of second-stage PCR reactions set up in triplicate. PCR reaction were set up as follows: 10 μ l of Buffer G, 0.2 μ l of Q5 polymerase (NEB), 0.8 μ l of i7 Nextera XT DNA Library Prep Kit primer (Illumina) and 0.4 μ l of an equimolar mixture of four reverse

primers (P5_BCodeNest_F1–4) mixed to a final concentration of 20 μ M. The PCR program used was 98 °C for 30 seconds, followed by 11 cycles of 98 °C for 20 seconds, 55 °C for 20 seconds, 72 °C for 45 seconds and a final extension of 72 °C for 60 seconds. The PCR reactions were then separated on a 1.5% agarose gel with the ~1,200-bp amplicon smear excised and purified. The amplicon was sequenced using MiSeq Reagent Nano Kit v2–300 cycles (Illumina). A custom Read 1 sequencing primer, Read1seq, was loaded into reagent cartridge position 18 as per the manufacturer's instructions.

PacBio sequencing processing

PacBio HiFi sequencing of the 10,000 metagenomic cosmids library generated 4.73 Gbp of data with an average insert length of 7 kb; the raw data were assembled using Flye version 2.9-b1768 in metagenomic mode (that is, metaFlye³⁵) with the 'pacbio-hifi' option. Generated contigs were analyzed by antiSMASH 5.1 (ref. ³⁶) for BGCs using the 'relaxed' setting and the '-cb-general-cb-knownclusters-cb-subclusters-genefinding-tool prodigal-m' options. For anti-phage mechanisms (that is, CRISPR–Cas), DefenseFinder²¹ was run using default settings. antiSMASH predicted 200 total clusters from which at least one example of each core biosynthetic feature was selected for recovery. To account for possible assembly or sequencing errors, contigs that contained multiple barcodes, as well as contigs with barcodes that did not match the expected degenerate sequence, were removed. A type 1 PKS cluster (contig 2,441) was predicted as two overlapping cosmids, and so both were recovered. In total, 70 cosmid targets accounting for 65 clusters and four CRISPR–Cas predictions were selected to demonstrate CCIC recovery. Sequences of the isolated cosmids have been deposited in GenBank, ON996267–ON996333, along with the barcode OTU lasso peptide BGC from edge mapping (GenBank OP058960).

Edge mapping

Read 1 sequences that contained barcodes were identified by searching for the fixed pattern 'CTAATTGGCCGTCGA' using the 'locate' command in SeqKit version 2.1.0 (ref. ³⁷). The 35-bp region upstream of this pattern, which contained the PAM, barcode, SpeI cloning site and an extra 5 bp, was extracted from each read. The extracted barcode sequences were then clustered at 94% identity using VSEARCH version 2.18.0 (ref. ³⁸) to generate OTUs. OTUs comprising at least ten sequences were carried forward, which gave 10,145 barcode OTUs. The paired Read 2 sequences associated with each barcode OTU were retrieved, trimmed to 100 bp and re-labeled to contain the barcode OTU ID (the script used is available at <https://doi.org/10.5281/zenodo.6574918>). The trimmed reads were then mapped to the *S. albidoflavus* J1074 reference genome (GenBank CP004370) using minimap2 version 2.24-r1122 (refs. ^{12,13}) and visualized using UGENE version 42.0 (ref. ³⁹). An example of the visualization is provided in Supplementary Fig. 5.

The *S. albidoflavus* reference genome was analyzed by antiSMASH (<https://antismash.secondarymetabolites.org>) with the default relaxed setting and all extra features selected. The antiSMASH output was used as an overview of the total BGC content as well as the boundaries of each cluster. Barcode OTUs associated with edge sequences that mapped close to a desired cluster were identified. Target barcode OTU sequences were retrieved, and a homologous sgRNA cloning primer was ordered for guide plasmid construction as described in the 'Cloning sgRNA constructs' methods section. The isolated region 8 cosmid contained a 17-kb segment of the genome beginning at the target OTU but was fused with 18 kb of a separate genomic region; the 17 kb was enough to capture the complete region 8 cluster.

Reporting summary

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

Data availability

pCCIC has been deposited as GenBank ON804120. Recovered BGC clusters containing cosmid clones have been deposited as GenBank ON996267–ON996333 and OP058960. The *S. albidoflavus* J1074 reference genome used is publicly available as GenBank CP004370. The Comprehensive Antibiotic Resistance Database, used for antibiotic resistance analysis of sucrose escape clones, is publicly available at <https://card.mcmaster.ca>. Other data are available from the corresponding authors upon reasonable request. Source data are provided with this paper.

Code availability

Publicly available Flye version 2.9-b1768, antiSMASH 5.1, Defense Finder 1.0.8, SeqKit version 2.1.0, VSEARCH version 2.18.0, minimap2 version 2.24-r1122 and UGENE version 42.0 were used for sequence assembly and analysis. Custom code generated for edge mapping has been deposited in Zenodo (<https://doi.org/10.5281/zenodo.6574918>).

References

- Sternberg, N., Ruether, J. & deRiel, K. Generation of a 50,000-member human DNA library with an average DNA insert size of 75–100 kbp in a bacteriophage P1 cloning vector. *New Biol.* **2**, 151–162 (1990).
- Zaburanyi, N., Rabyk, M., Ostash, B., Fedorenko, V. & Luzhetskyy, A. Insights into naturally minimised *Streptomyces albus* J1074 genome. *BMC Genomics* **15**, 97 (2014).
- Wu, C., Shang, Z., Lemetre, C., Ternei, M. A. & Brady, S. F. Cadasides, Calcium-dependent acidic lipopeptides from the soil metagenome that are active against multidrug-resistant bacteria. *J. Am. Chem. Soc.* **141**, 3910–3919 (2019).
- Jiang, Y. et al. Multigene editing in the *Escherichia coli* genome via the CRISPR–Cas9 system. *Appl. Environ. Microbiol.* **81**, 2506–2514 (2015).
- Chang, A. C. & Cohen, S. N. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J. Bacteriol.* **134**, 1141–1156 (1978).
- Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
- Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl Acad. Sci. USA* **70**, 3240–3244 (1973).
- Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
- Chan, W. T., Verma, C. S., Lane, D. P. & Gan, S. K. A comparison and optimization of methods and factors affecting the transformation of *Escherichia coli*. *Biosci. Rep.* **33**, e00086 (2013).
- Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
- Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
- Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
- Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
- Okonechnikov, K., Golosova, O., Fursov, M. & UGENE Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).

Acknowledgements

This work was supported by National Institutes of Health grant 5R35GM122559 (S.F.B.). We thank the Marraffini laboratory for pCas9.

Vectors pCas (Addgene, 62225) and pTargetF (Addgene, 62226) were gifts from Sheng Yang, and pwtCas9-bacteria (Addgene, 44250) was a gift from Stanley Qi. PacBio sequencing was performed by the Rockefeller University Vertebrate Genome Center.

Author contributions

J.B., V.K.L. and S.F.B. conceived CCIC retrieval. J.B. and S.F.B. designed and analyzed experiments. J.B. performed all experiments, with the aid of M.A.T. for metagenomic DNA preparation and lambda packaging, L.G. for various cloning and Y.A.H. for bioinformatics.

Competing interests

S.F.B. has consulted for Zymergen. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-022-01531-8>.

Correspondence and requests for materials should be addressed to Sean F. Brady.

Peer review information *Nature Biotechnology* thanks Benjamin Rubin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Custom code generated for Edge Mapping has been deposited in Zenodo (<https://doi.org/10.5281/zenodo.6574918>).

Data analysis Publicly available Flye v2.9-b1768, Antismash 5.1, Defense Finder 1.0.8, SeqKit v2.1.0, VSEARCH v2.18.0, minimap2 v2.24-r1122, and UGENE v42.0 were used for sequence assembly and analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

pCCIC has been deposited in Genbank ON804120, recovered BGC clusters containing cosmid clones have been deposited as Genbank ON996267-ON996333 and OP058960. The *S. albidoflavus* J1074 reference genome used is publicly available as Genbank CP004370. The Comprehensive Antibiotic Resistance Database (CARD),

used for antibiotic resistance analysis of sucrose escape clones, is publically available at <https://card.mcmaster.ca>. Other data are available from the corresponding authors upon reasonable request.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="text" value="n/a"/>
Population characteristics	<input type="text" value="n/a"/>
Recruitment	<input type="text" value="n/a"/>
Ethics oversight	<input type="text" value="n/a"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="No sample size calculations were performed. To establish CCIC for target retrieval, experiments were carried out in triplicate to account for any variance. CCIC retrieval from a metagenomic library (n=73) and a genomic library (n=12) established the robustness of the method."/>
Data exclusions	<input type="text" value="No data was excluded"/>
Replication	<input type="text" value="To establish CCIC retrieval parameters two independent guide RNA were used to ensure replication yielding comparable results. Eighty-five guides were used to show replication of retrieval of target clones from complex libraries with a 95% success rate."/>
Randomization	<input type="text" value="Randomization was not required as each experimental replicate was grown under identical conditions."/>
Blinding	<input type="text" value="Blinding was not relevant as each experimental replicate was grown under identical conditions."/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging