CRISPR-Cas9 enrichment, a new strategy in microbial metagenomics to investigate complex genomic regions: The case of an environmental integron

Eva Sandoval-Quintana, Christina Stangl, Lionel Huang, Ivo Renkens, Robert Duran, Gijs van Haaften, Glen Monroe, Béatrice Lauga, Christine Cagnon

MOLECULAR ECOLOGY
RESOURCES

CRISPR-Cas9 enrichment, a new strategy in microbial metagenomics to investigate complex genomic regions: the case of an environmental integron

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-22-0388
Manuscript Type:	Resource Article
Date Submitted by the Author:	08-Sep-2022
Complete List of Authors:	Sandoval-Quintana, Eva; Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA Stangl, Christina; University Medical Centre Utrecht, Department of Genetics Huang, Lionel; Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA Renkens, Ivo; University Medical Centre Utrecht, Department of Genetics Duran, Robert; Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA van Haaften, Gijs; University Medical Centre Utrecht, Department of Genetics Monroe , Glen; University Medical Centre Utrecht, Department of Genetics Lauga, Béatrice; Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA Cagnon, Christine; Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA
Keywords:	complex genomic regions, CRISPR-Cas9 enrichment, environmental integrons, microbial communities, microbial metagenomics, mobile genetic elements

CRISPR-CAS9 ENRICHMENT, A NEW STRATEGY IN MICROBIAL METAGENOMICS TO INVESTIGATE COMPLEX GENOMIC REGIONS: THE CASE OF AN ENVIRONMENTAL INTEGRON

Running title

CRISPR-Cas9 enrichment for metagenomics

Eva Sandoval-Quintana[1], Christina Stangl[2], Lionel Huang[1], Ivo Renkens[2], Robert Duran[1], Gijs van Haaften[2], Glen Monroe[2], Béatrice Lauga[1,*,§], Christine Cagnon[1,*,§].

[1] Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, MIRA, Pau, France

[2] Department of Genetics, University Medical Center Utrecht, Utrecht, Netherlands.

* Correspondence: christine.cagnon@univ-pau.fr (C. Cagnon); beatrice.lauga@univ-pau.fr (B. Lauga)

§ Co-last authors

Abstract

Environmental integrons are ubiquitous in natural microbial communities, but they are mostly uncharacterized and their role remains elusive. Thus far, research has been hindered by methodological limitations. Here, we successfully used an innovative approach combining CRISPR-Cas9 enrichment with long-read nanopore sequencing to target, in a complex microbial community, a putative adaptive environmental integron, InOPS, and to unravel its complete structure and genetic context. A contig of 20 kb was recovered containing the complete integron from the microbial metagenome of oil-contaminated coastal sediments. InOPS exhibited typical integron features. The integrase, close to integrases of marine Desulfobacterota, possessed all the elements of a functional integron integrase. The gene cassettes harboured mostly unknown functions hampering inferences about their ecological importance. Moreover, the putative InOPS host, likely a hydrocarbonoclastic marine bacteria, might question the adaptive potential of InOPS in response to oil contamination. Additionally, several mobile genetic elements were intertwined with InOPS highlighting likely genomic plasticity, and providing a source of genetic novelty. This case study showed the power of CRISPR-Cas9 enrichment to elucidate the structure and context of specific DNA regions for which only a short sequence is known. This method is a new tool for environmental microbiologists working with complex

microbial communities to target low abundant, large or repetitive genetic structures that are hard to recover by classical metagenomics. More precisely, here, it offers new perspectives to comprehensively assess the eco-evolutionary significance of environmental integrons.

Keywords

complex genomic regions, CRISPR-Cas9 enrichment, environmental integrons, microbial communities, microbial metagenomics, mobile genetic elements

1. Introduction

Integrons are genetic elements that acquire, excise, rearrange and express gene cassettes (Fig. 1) (Escudero *et al.,* 2015; Stokes & Hall, 1989). They are notorious for mediating bacterial adaptation in clinical settings through the spread of antibiotic resistance genes (Escudero *et al.,* 2015; Gillings, 2014; Souque *et al.,* 2021). Several clues suggest that integrons in the

environmental may also participate in the bacterial functional response to selective pressures, *e.g.,* their occurrence in numerous environments (e.g., Abella, Bielen, *et al.,* 2015; Abella, Fahy, *et al.,* 2015; Antelo *et al.,* 2021; Boucher *et al.,* 2007; Elsaied, Stokes, Nakamura *et al.,* 2007; Elsaied, Stokes, Kitamura *et al.,* 2011; Elsaied, Stokes, Yoshioka *et al.,* 2014; Ghaly, Penesyan *et al.,* 2022; Mazel, 2006; Nemergut *et al.,* 2004; Stokes *et al.,* 2001), distribution in diverse bacterial taxa (Cambray *et al.,* 2010; Cury *et al.,* 2016), wide gene cassette functional repertory (e.g., Ghaly, Geoghegan *et al.,* 2019; Pereira *et al.,* 2020), and relation to environmental disturbances (Abella, Fahy, *et al.,* 2015; Elsaied, Stokes, Kitamura *et al.,* 2011; Koenig, Boucher *et al.,* 2008; Koenig, Sharp *et al.,* 2009; Nemergut *et al.,* 2004). Nonetheless, in contrast to clinical integrons, environmental integrons remain mostly uncharacterized and their role is still elusive (Sandoval-Quintana *et al.,* 2022).

Establishing their adaptive role in the environment requires the characterization of their complete genetic structures including the integrase gene, gene cassettes, recombination sites, and the genetic context, to reveal the function of their gene cassettes, their functionality, their dynamics and eventually, their host. However, targeting putative adaptive integrons and their complete platforms in complex communities is a difficult task hindered by methodological limitations (Pereira *et al.,* 2020; Sandoval-Quintana *et al.,* 2022). To date, integrons in natural environments have been mainly recovered through amplicon sequencing (e.g., Abella, Bielen

et al., 2015; Antelo *et al.,* 2021; Ghaly, Geoghegan *et al.,* 2019; Ghaly, Penesyan *et al.,* 2022). Because of the constraints imposed by the design of the primers located in the conserved and known regions of integrases and recombination sites (Elsaied, Stokes, Nakamura *et al.,* 2007; Stokes *et al.,* 2001) and the impediments in amplifying large fragments, efforts have until recently mainly focused either on gene cassettes (e.g., Ghaly, Geoghegan *et al.,* 2019; Koenig, Boucher *et al.,* 2008; Koenig, Sharp *et al.,* 2009; Stokes *et al.,* 2001) or integrase genes (e.g. Abella, Bielen *et al.,* 2015; Elsaied, Stokes, Nakamura *et al.,* 2007; Nield *et al.,* 2001), preventing the comprehensive examination of entire environmental integron platforms.

Although these efforts did reveal the extent of their diversity, their dynamics (integration, excision events, dispersal) were poorly assessed, despite attempts to infer ongoing integron-response to selective pressures being developed (Huang *et al.,* 2009). Overall, however, all these PCR-based strategies are still impaired by the well-known PCR bias, the inability to position gene cassettes in the array, and the difficulties of designing universal primers to target environmental integrons. Together with PCR approaches, DNA/DNA hybridization screening techniques were also used to recover novel environmental integrases from large metagenomic clone libraries (Jacquiod *et al.,* 2014, Moura, Henriques *et al.,* 2010). Nevertheless, this approach can suffer from parasite signals, unspecific probe hybridization, and low target recovery (Jacquiod *et al.,* 2014; Moura, Henriques *et al.,* 2010).

Whole genome and metagenome sequencing and subsequent analyses with adequate bioinformatic tools offer the opportunity to mine for integrons in isolated strains and in uncultured microorganisms, respectively (Cury *et al.,* 2016; Pereira *et al.,* 2020). Nevertheless, the highly fragmented nature of shotgun sequence data and the frequent association of integrons to repetitive regions (Pereira *et al.,* 2020; Sonbol & Siam, 2021) have often impeded the assembly of large contigs, and thus, the recovery of full-length integrons preventing further conclusions about their ecological-evolutionary significance.

The main objective of our study was to propose a new methodology to recover the complete platforms of putative adaptive integrons to investigate their role in the environment. The first step of the study consisted in the identification of a putative adaptive integron in a complex microbial community. We thus experimentally mimicked an environmental disturbance, an oil contamination, on coastal sediments maintained in mesocosms. We then proposed a new approach based on CRISPR-Cas9 enrichment to recover the InOPS environmental integron from the complex microbial metagenome. This method allowed successfully to decipher the complete structure of InOPS, highlighting its relevance to assess the eco-evolutionary significance of this integron.

2. Material and methods

2.1. Coastal sediment samples

Coastal marine sediments (Atlantic Ocean, Brittany, France) were maintained during 9 months in mesocosms as previously described (Stauffert *et al.,* 2013). Mesocosms were contaminated by oil addition (OIL). Control mesocosms (CTRL) were maintained without contamination. Samples (0 to 2 cm depth) were collected at different times: just after the oil contamination (T0), after one month, six months, and nine months.

2.2. Identification and quantification of intIOPS

Metagenomic DNA was extracted with Ultraclean soil DNA isolation kit (MoBio Laboratories, Carlsbad, CA, USA). Integron integrase gene (*intI)* libraries of approximately 770 bp fragments were obtained from mesocosms incubated during 9 months under oil contamination or no contamination as previously described (Abella, Fahy, *et al.,* 2015). Sanger sequencing was performed by GATC Biotech (Ebersberg, Germany). Sequences were corrected with Sequencher® (Accession numbers: FR718193-FR718248), then were clustered at 100% identity using CD-HIT (Fu *et al.,* 2012) before calculating Shannon (Shannon, 1948), evenness (Pielou, 1966), and coverage (Good, 1953) indexes.

Quantitative PCRs were performed in duplicate on metagenomic DNA from mesocosms. DNA was quantified with Quant-iT™ PicoGreen® dsDNA (Invitrogen, Waltham, MA, USA).

Standard curves, constructed with plasmid carrying cloned genes, were used to quantify *intIOPS* and 16S rRNA genes, respectively. The primers ICC60 (5'-GAAACCGTTCGGTTGAGGGTC-3') and ICC71 (5'-TTTACGGGCCAGCGCACCGGG-3') were used to specifically amplify *intIOPS*. The primers 338F and 518R (Lane, 1991) were used to amplify the 16S rRNA genes. The reaction mixture contained 1 µL of template DNA, 0.2 µM of each primer and 12.5 µL of Power SYBR® green PCR master mix (Applied Biosystem, Waltham, MA, USA) for a final volume of 25 µL. All real-time PCRs were performed on a MX3005P (Stratagene, San Diego, CA, USA). Amplification parameters were as follows: 10 min at 95°C, 40 cycles of 30 s at 95°C, 30 s at 56°C (*intIOPS* gene) or 55°C (16S rRNA gene), and 45 s at 72°C. After amplification, a melting curve was carried out to confirm the amplification of specific products.

2.3. CRISPR-Cas9 targeted enrichment of InOPS from metagenomic DNA coupled to nanopore sequencing

High molecular weight (HMW) DNA was obtained using the DNeasy PowerSoil Pro Kit (QiAgen, Hilden, Germany) with slight modifications to prevent DNA shearing. DNA quantification was performed using the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). DNA quality and integrity were assessed through standard absorbance ratios and using a 4200 TapeStation System (Agilent, Santa Clara, CA, USA), respectively.

CRISPR RNA (crRNA) specific probes were designed to target *intIOPS* within the integron integrase additional domain region (Messier & Roy, 2001), according to several criteria, including probe directionality (here, upstream or downstream the region of interest), downstream presence of a protospacer adjacent motif (PAM) site (5' - NGG – 3'), and on-target efficiency score (Stangl *et al.,* 2020). crRNA were designed mainly using the *IDT Custom Alt-R® CRISPR-Cas9 gRNA tool* (IDT, Coralville, IA, USA) and additionally *CHOPCHOP* (Labun *et al.,* 2019) and *CRISPOR* (Concordet & Haeussler, 2018). As working with metagenomes, no reference genome and no off-target values were considered. One crRNA (5'-AAAAAGGCGGAAAAGAGCTG -3') was designed to uncover the unknown 5' part of the integron, expected to be its genetic context. The other crRNA (5'-TCCACCGACAAGGTTTTGGA -3') was designed to uncover the unknown 3' part of the integron, expected to be the gene cassette array. Custom Alt-R® crRNAs were ordered from Integrated DNA Technologies (IDT, Coralville, IA, USA).

Cleavage of a 378 bp amplicon of the target DNA obtained with ICC66 (5'-GATCGACAACCATGGGGGGAG-3') and ICC67 (5'-TGGTGACGCCGCTTGACACC-3') primers was performed to validate the efficiency of the crRNAs prior to metagenomic enrichment. Digestion was performed according to IDT recommendation (*Alt-R CRISPR-Cas9 system—in vitro cleavage of target DNA with RNP complex* protocol) with slight modifications

(excess of ribonucleoprotein complexes (RNPs) for digestion, 20 min digestion and addition of a final step at 72°C for 5 min to inactivate proteinase K and obtain the complete Cas9 release) and checked on a 4200 TapeStation System (Agilent).

CRISPR-Cas9 mediated enrichment of InOPS from HMW metagenomic DNA was performed following the protocol of (Stangl *et al.,* 2020). Two enrichments were independently performed, one with the crRNA targeting the genetic context and the other with the crRNA targeting the gene cassette array. In brief, approximately 4.3 µg of HMW metagenomic DNA were dephosphorylated. crRNA was annealed to the trans-activating CRISPR RNA (tracrRNA) and then, RNP were formed by adding HiFi Cas9 enzyme (IDT, Coralville, IA, USA). The Cas9 RNP were mixed with the dephosphorylated DNA, dATP and Taq polymerase to produce the targeted double-strand breaks and facilitate dA-tailing. Oxford Nanopore Technologies (ONT) specific sequencing adapters (SQK-LSK109, ONT, Oxford, UK) were ligated to the free phosphorylated ends. Libraries were purified with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) with fragments below 3 kb washed away. The DNA concentration of the enriched libraries was measured using the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific). The two libraries were pooled prior to sequencing on a single ONT flow cell (R9.4, ONT, Oxford, UK) according to the manufacturer's protocol. Sequencing was performed on a

GridION X5 instrument (ONT, Oxford, UK; Utrecht Sequencing Facility, Utrecht, The Netherlands).

2.4. Characterization of the InOPS contig

Base-calling of nanopore reads was performed by Guppy (ONT, Oxford, UK) with the high accuracy model (Q-score cut-off >7). Sequencing library statistics were generated using Nanoplot (v 1.28.2) (De Coster *et al.,* 2018). ONT adapters were trimmed off using Porechop (v. 0.2.4) (Wick, 2018) with default parameters. Reads were mapped to the *intIOPS* reference sequence (FR718193.1) using primarily minimap2 (-x -map-ont) (v. 2.6) (Li, 2018), then BWA-MEM (-k:12, -O:4, -L: 5, -B: 1, -U:12) (v. 0.7.17) (Li & Durbin, 2010) and NCBI BLAST+ blastn (v. 2.10.1) (Camacho *et al.,* 2009) with default parameters for accuracy. Files were sorted and indexed with Samtools (v. 1.4.1) (Li *et al.,* 2009) and Bamtools (v. 2.4.0) (Barnett *et al.,* 2011). Sequences were aligned to the *intIOPS* reference sequence with the MUSCLE Multiple Alignment tool (Edgar, 2004). A manually curated consensus was created. This contig (InOPS contig) bearing the InOPS integron was validated by Sanger sequencing.

The contig was annotated using different annotation tools and pipelines to gain as much accuracy as possible. The open reading frames (ORFs) and coding sequences were predicted using Prokka (--meta and --evalue 1e-06) (v. 1.14.6) (Seemann, 2014), DFAST (--Metageannotator) (v. 1.5.0) (Tanizawa *et al.,* 2018), RASTtk (Brettin *et al.,* 2015), Contig

Annotation Tool (CAT_prepare_20210107 and --add_names) (v. 5.0.3) (von Meijenfeldt *et al.,* 2019), MetaGeneAnnotator (--meta option) (Noguchi *et al.,* 2008), MetaGeneMark (v. 3.25) (Zhu *et al.,* 2010) and/or GeneMarkS (Besemer *et al.,* 2001)/S-2 (Lomsadze *et al.,* 2018). Integron Finder (v. 1.5.1) (Cury *et al.,* 2016) was used for integron detection (--local-max --evalue-attc 1) and *attC* sites detection (--local-max --evalue-attc 4 and --dt 6000). *attC* secondary structure was predicted using *mfold* (UNAFold) (Markham & Zuker, 2008) and the RNAfold program (ViennaRNA Package) (-- p –d2) (Lorenz *et al.,* 2011). The *attI* site was searched manually and aligned against an *attI* site database constructed from (Collis & Hall, 2004; Elsaied, Stokes, Kitamura *et al.,* 2011; Nield *et al.,* 2001; Partridge *et al.,* 2000). Promoters were identified using BPROM (Solovyev & Salamov, 2011) and PePPER (de Jong *et al.,* 2012) and further manually curated. The putative functionality of the annotated genes was inferred through Prokka, DFAST, RASTtk, EggNOG-mapper (--evalue 0.001 --itype metagenome --genepred prodigal --pfam_realign none) (v. 2.1.6) (Cantalapiedra *et al.,* 2021), InterProScan (--pathways --goterms) (v. 5.54-87.0) (Jones *et al.,* 2014) and HMMER (--hmmscan) (v. 2.41.2) (Potter *et al.,* 2018) against different databases (NCBI, SEED, Clusters of Orthologs Groups (COGs), Pfam, SMART, TIGRFAM, SFLD, SUPERFAMILY, PANTHER, Gene3d, HAMAP, PROSITE, Coils, PRINTS, PIRSR, PIRSF). The genes annotated as putative gene cassettes were confronted to the INTEGRALL database (Moura, Soares *et al.,*

209 2009) using local BLASTn algorithm. Kraken2 (v. 2.1.1) (Wood *et al.,* 2019) and CAT were
210 used to perform taxonomic classification of the annotated genes.

211 Insertion sequences (IS) were annotated using ISsaga (v. 2.0) (Varani *et al.,* 2011) and
212 OASIS (Robinson *et al.,* 2012). The putative insertion sequences (IS) were further analysed
213 with the ISFinder (Siguier *et al.,* 2006) BLAST interface. IS putative ORFs were compared
214 against local IS91 and ISCR databases through local Blastp algorithm. Group II introns were
215 identified against the bacterial Group II intron database (Candales *et al.,* 2012).

216 For synteny analysis, the annotated genes within the InOPS contig were compared to
217 databases through BLAST searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi): *blastn* against
218 the *nr/nt* database and the *wgs* database (*Desulfobacterales* taxid: 23118); *blastx* and *blastp*
219 against *nr/nt* and *env_nr* databases (NCBI). Resultant genomes (> 70% identity and 50%
220 coverage) were downloaded from NCBI constituting a database of 76 genomes after
221 dereplication. The genomic dataset was submitted to M1CR0B1AL1Z3R (Avram *et al.,* 2019)
222 along with the InOPS contig. Two different ortholog detections (≥ 80 or 50% identity and 0.01
223 as maximal *e-value*) were performed. To refine the ortholog detection, a second
224 M1CR0B1AL1Z3R was run over the genomes which produced a hit in the first run.
225 SimpleSynteny (Veltri *et al.,* 2016) was used for visualization with parameters in regular mode
226 (1 *e-value* and 25% coverage) using the *all-to-all comparison* mode.

2.5. Conservation and phylogeny of the InOPS integrase

InOPS integrase (IntIOPS) was compared against available databases (INTEGRALL (Moura, Soares *et al.,* 2009) and NCBI) using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) in *blastn, blastx* and *blastp* mode. IntIOPS, IntI1-4 (AHL30833.1, AAT72891.1, AAO32355.1, AAC38424.1) whose functionality has been largely studied, IntIs issued from natural environments (IntINeu: WP_011112687.1, InPstQ: AAN16061.1, SamIntIA: WP_011759470.1, IntIPac: AAK73287.1, IntISon: WP_011072111.1) whose activity has been experimentally proved and the *Escherichia coli* XerD recombinase (P0A8P8.1) were aligned using MUSCLE Multiple Alignment tool (Edgar, 2004). The alignment was manually edited with BioEdit® software.

For the IntI tree, a dataset was built by selecting 128 amino acid sequences of complete integron integrases (IntI). Sequences were retrieved from *nr/nt* and *env_nr* databases (NCBI) using both *blastx* and *blastp* and selected based on their identity to the integrase of the InOPS integron, IntIOPS (≥ 50% identity and coverage). To avoid redundancy, the sequences were clustered to 90% identity using CD-HIT. Within each cluster, the representative sequence was kept and, considering its environmental origin, the closest sequence from each different environmental origin, if any, were identified and kept too. Sequences of clinical IntIs, IntI1-IntI4 (AHL30833.1, AAT72891.1, AAO32355.1, AAC38424.1), and the integron integrases

belonging to the genus *Desulfosarcina* (WP_051374975.1, WP_027353082.1, BBO66607.1, BBO93164.1, BBO87010.1, BBO79603.1, WP_083456647.1, WP_198012316.1, WP_155322972.1) were also included. The final dataset comprised 107 integron integrase sequences. The tyrosine recombinase XerD sequence of *Escherichia coli* (P0A8P8.1) was included as outgroup for the construction of the tree. Analysis were done using NGPhylogeny.fr (Lemoine *et al.,* 2019) with the options of MAFFT align and BMGE alignment curation. The PhyML program (Guindon *et al.,* 2010) was used for tree construction with the SMS option (Lefort *et al.,* 2017) and a bootstrapping branch support of 1000. The tree was submitted to iTOL (Letunic & Bork, 2021) for visualization and design.

3. Results

3.1. Mimicking environmental disturbance identified InOPS as a putative adaptive integrons

We investigated coastal sediments exposed to oil contamination in mesocosms to trigger microbial community adaptive response (Stauffert *et al.,* 2013). We revealed, using PCR targeting integron integrase genes (*intI*) (Abella, Bielen, *et al.,* 2015), the predominance of a sequence, named *intIOPS*. This sequence represented nearly 27% (over 156 amplicon sequences) of the *intI* pool in the amplicon libraries generated from the mesocosms incubated under oil contamination while no *intIOPS* sequences (over 46 amplicon sequences) were

263 detected in the library generated from the control (without contamination). Diversity indexes
264 supported the lower diversity and divergent relative abundance of *intI* in the contaminated
265 sediments compared to the control (Shannon: 3.14 and 3.41 *vs* 3.61, Evenness: 0.82 and 0.87
266 *vs* 0.98). The increase of *intIOPS* after contamination was further supported using quantitative
267 PCR (Fig. S1). Therefore, *intIOPS* might belong to an environmental integron responding to
268 the oil contamination. We named this integron InOPS.

269 **3.2. Based FUDGE CRISPR-Cas9 enrichment to target *intIOPS* integrase gene**

270 The InOPS integron was recovered from the complex microbial metagenome by an
271 innovative approach derived from FUDGE (Stangl *et al.,* 2020), a CRISPR-Cas9 enrichment
272 method coupled to nanopore sequencing, that only targets a short-conserved sequence (Fig.
273 2A for full description). Two crRNAs were designed within a 62 bp region of the *intIOPS*
274 additional domain to obtain both flanking unknown regions (Fig. 2A, 2B). Sequencing of the
275 enriched nanopore libraries representing 1.21 Gb of sequence data resulted in 546 194 good
276 quality reads (N50: 3 337 bp, average read length: 1 453.6 bp). We recovered 90 reads
277 towards the 3' unknown part of the integron containing the array of gene cassettes (N50: 4 258
278 bp, longest read: 13 201 bp) and 51 reads towards the 5' unknown part of the integron
279 corresponding to the genetic context (N50: 4 146 bp, longest read: 8 606 bp) (Fig. 2B). Overall,

0.03% of the reads covered the targeted sequences. The consensus contig (20 069 bp) was further polished using Sanger sequencing.

3.3. Unraveling the InOPS full integron platform structure and its genetic context

The contig annotation showed that the complete InOPS integron was recovered. It exhibited the typical integron features (Sandoval-Quintana *et al.,* 2022): an integron integrase gene, putative functional *attI* site, PintI and Pc promoters and regulator binding sites, as well a gene cassette array (Fig. 3; Fig. S2; Tables S1 and S2). The integrase possessed the catalytic residues and most of the conserved motifs of integron integrases (Messier & Roy, 2001) (Fig. S3) suggesting the enzyme is functional. It only presented 72% identity to its closest relative and was divergent from the clinical integron integrase classes (≤ 50% identity). InOPS integrase clustered with integron integrases issued from environmental sources, in a consistent manner from marine environments from which InOPS originated, and with integron integrases belonging to Desulfobacterota (Fig. S4).

The cassette array consists of 12 gene cassettes with their own *attC* recombination site. Variable in length and sequences, 8 *attC* presented the typical secondary structure of *attC* sites suggesting their possible recombinogenic activity (Fig. S5). Most gene cassettes (apart the first one) encoded unknown functions or were ORFans, while others exhibited conserved

297 domains without obvious relationships with the oil contamination. Moreover, none were
298 referenced as gene cassettes in the INTEGRALL database (Moura, Soares *et al.,* 2009).

299 Interestingly, several mobile genetic elements (MGEs) were intertwined with InOPS. The
300 first gene cassette contained a complete IS of the IS91 family. IS91 can mobilize adjacent DNA
301 sequences and therefore participate in genomic plasticity (Garcillán-Barcia & de la Cruz,
302 2002). Here, it might disseminate InOPS elements. The 5' InOPS genetic context harbored a
303 complete IS1634 as well as other putative IS-like elements (Table S3). Such configurations
304 have been previously described (Cury *et al.,* 2016; Huyan *et al.,* 2020). A putative CALIN
305 embedded within this IS-rich region was also consistent with their frequent association with IS
306 (Cury *et al.,* 2016). Additionally, reverse transcriptase and maturase domains of a putative
307 group IIB intron were identified in the 3' InOPS genetic context. Of note, group II introns have
308 previously been observed associated with integrons (Léon & Roy, 2009; Sonbol & Siam, 2021),
309 and in some cases, hypothesized to be implicated in the genesis of gene cassettes (Léon &
310 Roy, 2009).

311 **3.4. Deciphering the origin of InOPS**

312 The lack of synteny evidenced that the configuration of ORFs within the contig is unique.
313 Variation in GC content over the contig clearly distinguished the gene cassettes (except IS91)
314 from the genetic context (Fig. 3), suggesting a different origin. The genetic context gave clues

about the InOPS host, likely belonging to Desulfobacterales (Table S1) and more precisely to *Desulfosarcina ovata,* a sulfate-reducing hydrocarbon-degrading and marine bacteria (Watanabe *et al.,* 2020). However, InOPS integrase divergence from *Desulfosarcina* integrases suggests the acquisition of InOPS functional platform from another Desulfobacterota (Fig. S4).

4. Discussion

With CRISPR-Cas9 enrichment we propose a new strategy in microbial metagenomics to capture large and specific regions in a simple way, without DNA amplification, with minimum required information, and avoiding time-consuming and haphazard metagenomic mining. While all our previous attempts failed, thanks to CRISPR-Cas9 enrichment, we retrieved and reliably deciphered the complete structure of the InOPS environmental integron. InOPS is an example of low abundant metagenomic regions but also complex, harboring repetitive sequences that jeopardize the use of standard metagenomic approaches.

Although the percentage of reads targeting the region we wished to study appeared low, the reads generated was abundant enough to retrieve InOPS. This case study demonstrated the efficiency of CRISPR-Cas9 enrichment in microbial metagenomics. It allowed to reach a level of resolution rarely equaled in the study of environmental integrons. Most gene cassettes

encoded unknown functions or were ORFans, a common feature for environmental integrons (Pereira *et al.,* 2020), raising more broadly the question about the origin of gene cassettes of integrons. Thus, the functions of InOPS gene cassettes are mostly unresolved, precluding conclusions about their ecological importance. Because we inferred the InOPS host, likely a hydrocarbonoclastic marine bacteria, the adaptive potential of InOPS facing oil contamination remains questionable. Tight association of InOPS with MGEs highlight that this region is subjected to genomic plasticity, as previously suggested for integrons in hypersaline environments (Sonbol & Siam, 2021), and might promote genetic novelty.

CRISPR-Cas9 enrichment offers the opportunity to reconsider studies that have previously identified adaptive gene cassettes (e.g., Elsaied, Stokes, Yoshioka *et al.,* 2014; Koenig, Sharp *et al.,* 2009; Nemergut *et al.,* 2004) or environmental integron integrases (e.g., Abella, Fahy, *et al.,* 2015; Elsaied, Stokes, Nakamura *et al.,* 2007; Nield *et al.,* 2001). Compiling such case studies, complemented with further molecular investigation on functionality and dynamics of integrons, could serve as a lever to assess the eco-evolutionary significance of environmental integrons. For instance, it could be of interest to decipher the interplay of environmental integrons with MGEs more comprehensively.

Overall, InOPS constitute a proof of concept that opens perspectives to document the dark matter of metagenomes and for which little information is available.

Acknowledgements

Research on environmental integrons was supported by funds from E2S-UPPA programs to CC (Initiative program and Innovation Research MIRA program) and BL (Hub-MeSMic project), and the ANR/SEST (06SEST09 and ANR450 CESA-2011-006 01 projects). ESQ was supported by PhD and mobility grants from E2S-UPPA programs and an EMBO Short-Term fellowship (STF-8420). LH was supported by a PhD grant from the Ministère de l'Enseignement Supérieur et de la Recherche (France). The authors thank Ophélie Tramoni for technical assistance. We thank Utrecht Sequencing Facility for providing sequencing service and data. Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute,Utrecht University and The Netherlands X-omics Initiative (NWO project 184.034.019).

References

Abella, J., Bielen, A., Huang, L., Delmont, T. O., Vujaklija, D., Duran, R., & Cagnon, C. (2015). Integron diversity in marine environments. *Environmental Science and Pollution Research*, *22*(20), 15360–15369. https://doi.org/10.1007/s11356-015-5085-3

Abella, J., Fahy, A., Duran, R., & Cagnon, C. (2015). Integron diversity in bacterial communities of freshwater sediments at different contamination levels. *FEMS Microbiology Ecology*, *91*(12), fiv140. https://doi.org/10.1093/femsec/fiv140

Antelo, V., Giménez, M., Azziz, G., Valdespino-Castillo, P., Falcón, L. I., Ruberto, L. A. M., Mac Cormack, W. P., Mazel, D., & Batista, S. (2021). Metagenomic strategies identify diverse integron-integrase and antibiotic resistance genes in the Antarctic environment. *MicrobiologyOpen*, *10*(5), e1219. https://doi.org/10.1002/mbo3.1219

Avram, O., Rapoport, D., Portugez, S., & Pupko, T. (2019). M1CR0B1AL1Z3R—a user-friendly web server for the analysis of large-scale microbial genomics data. *Nucleic Acids Research*, *47*(W1), W88–W92. https://doi.org/10.1093/nar/gkz423

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, *27*(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*(12), 2607–2618. https://doi.org/10.1093/nar/29.12.2607

Boucher, Y., Labbate, M., Koenig, J. E., & Stokes, H. W. (2007). Integrons: Mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology*, *15*(7), 301–309. https://doi.org/10.1016/j.tim.2007.05.004

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*, 8365. https://doi.org/10.1038/srep08365

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cambray, G., Guerout, A.-M., & Mazel, D. (2010). Integrons. *Annual Review of Genetics*, *44*, 141–166. https://doi.org/10.1146/annurev-genet-102209-163504

Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., McNeil, B. A., Wu, L., Jarding, A. M., & Zimmerly, S. (2012). Database for bacterial group II introns. *Nucleic Acids Research*, *40*(D1), D187–D190. https://doi.org/10.1093/nar/gkr1043

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Collis, C. M., & Hall, R. M. (2004). Comparison of the structure-activity relationships of the integron-associated recombination sites attI3 and attI1 reveals common features.

Microbiology (Reading, England), *150*(Pt 5), 1591–1601. https://doi.org/10.1099/mic.0.26596-0

Concordet, J.-P., & Haeussler, M. (2018). CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, *46*(W1), W242–W245. https://doi.org/10.1093/nar/gky354

Cury, J., Jové, T., Touchon, M., Néron, B., & Rocha, E. P. (2016). Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Research*, *44*(10), 4539–4550. https://doi.org/10.1093/nar/gkw319

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P., & Kok, J. (2012). PePPER: A webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics*, *13*(1), 299. https://doi.org/10.1186/1471-2164-13-299

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Elsaied, H., Stokes, H. W., Kitamura, K., Kurusu, Y., Kamagata, Y., & Maruyama, A. (2011). Marine integrons containing novel integrase genes, attachment sites, attI , and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *The ISME Journal*, *5*(7), 1162–1177. https://doi.org/10.1038/ismej.2010.208

Elsaied, H., Stokes, H. W., Nakamura, T., Kitamura, K., Fuse, H., & Maruyama, A. (2007). Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environmental Microbiology*, *9*(9), 2298–2312. https://doi.org/10.1111/j.1462-2920.2007.01344.x

Elsaied, H., Stokes, H. W., Yoshioka, H., Mitani, Y., & Maruyama, A. (2014). Novel integrons and gene cassettes from a Cascadian submarine gas-hydrate-bearing core. *FEMS Microbiology Ecology*, *87*(2), 343–356. https://doi.org/10.1111/1574-6941.12227

Escudero, J. A., Loot, C., Nivina, A., & Mazel, D. (2015). The Integron: Adaptation on demand. *Microbiology Spectrum*, *3*(2), 1–22, MDNA3-0019–2014. https://doi.org/10.1128/microbiolspec.MDNA3-0019-2014

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Garcillán-Barcia, M. P., & de la Cruz, F. (2002). Distribution of IS91 family insertion sequences in bacterial genomes: Evolutionary implications. *FEMS Microbiology Ecology*, *42*(2), 303–313. https://doi.org/10.1111/j.1574-6941.2002.tb01020.x

Ghaly, T. M., Geoghegan, J. L., Alroy, J., & Gillings, M. R. (2019). High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environmental Microbiology*, *21*(5), 1567–1574. https://doi.org/10.1111/1462-2920.14551

Ghaly, T. M., Penesyan, A., Pritchard, A., Qi, Q., Rajabal, V., Tetu, S. G., & Gillings, M. R. Y. 2022. (2022). Methods for the targeted sequencing and analysis of integrons and their gene cassettes from complex microbial communities. *Microbial Genomics*, *8*(3), 000788. https://doi.org/10.1099/mgen.0.000788

Gillings, M. R. (2014). Integrons: Past, present, and future. *Microbiology and Molecular Biology Reviews: MMBR*, *78*(2), 257–277. https://doi.org/10.1128/MMBR.00056-13

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Huang, L., Cagnon, C., Caumette, P., & Duran, R. (2009). First gene cassettes of integrons as targets in finding adaptive genes in metagenomes. *Applied and Environmental Microbiology*, *75*(11), 3823–3825. https://doi.org/10.1128/AEM.02394-08

Huyan, J., Tian, Z., Zhang, Y., Zhang, H., Shi, Y., Gillings, M. R., & Yang, M. (2020). Dynamics of class 1 integrons in aerobic biofilm reactors spiked with antibiotics. *Environment International*, *140*, 105816. https://doi.org/10.1016/j.envint.2020.105816

Jacquiod, S., Demanèche, S., Franqueville, L., Ausec, L., Xu, Z., Delmont, T. O., Dunon, V., Cagnon, C., Mandic-Mulec, I., Vogel, T. M., & Simonet, P. (2014). Characterization of new bacterial catabolic genes and mobile genetic elements by high throughput genetic screening of a soil metagenomic library. *Journal of Biotechnology*, *190*, 18–29. https://doi.org/10.1016/j.jbiotec.2014.03.036

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Koenig, J. E., Boucher, Y., Charlebois, R. L., Nesbø, C., Zhaxybayeva, O., Bapteste, E., Spencer, M., Joss, M. J., Stokes, H. W., & Doolittle, W. F. (2008). Integron-associated

gene cassettes in Halifax Harbour: Assessment of a mobile gene pool in marine sediments. *Environmental Microbiology*, *10*(4), 1024–1038. https://doi.org/10.1111/j.1462-2920.2007.01524.x

Koenig, J. E., Sharp, C., Dlutek, M., Curtis, B., Joss, M., Boucher, Y., & Doolittle, W. F. (2009). Integron gene cassettes and degradation of compounds associated with industrial waste: The case of the Sydney tar ponds. *PLOS ONE*, *4*(4), e5276. https://doi.org/10.1371/journal.pone.0005276

Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., & Valen, E. (2019). CHOPCHOP v3: Expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research*, *47*(W1), W171–W174. https://doi.org/10.1093/nar/gkz365

Lane, D. J. (1991). 16S/23S rRNA sequencing. *Stackebrandt, E. and Goodfellow, M., Eds., Nucleic Acid Techniques in Bacterial Systematic*, 115–175.

Lefort, V., Longueville, J.-E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, *34*(9), 2422–2424. https://doi.org/10.1093/molbev/msx149

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., & Gascuel, O. (2019). NGPhylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Research*, *47*(W1), W260–W265. https://doi.org/10.1093/nar/gkz303

Léon, G., & Roy, P. H. (2009). Potential Role of Group IIC-attC Introns in Integron Cassette Formation. *Journal of Bacteriology*, *191*(19), 6040–6051. https://doi.org/10.1128/JB.00674-09

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lomsadze, A., Gemayel, K., Tang, S., & Borodovsky, M. (2018). Modeling leaderless transcription and atypical genes results in more accurate gene prediction in

prokaryotes. *Genome Research*, *28*(7), 1079–1089. https://doi.org/10.1101/gr.230615.117

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Markham, N. R., & Zuker, M. (2008). UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular Biology (Clifton, N.J.)*, *453*, 3–31. https://doi.org/10.1007/978-1-60327-429-6_1

Mazel, D. (2006). Integrons: Agents of bacterial evolution. *Nature Reviews. Microbiology*, *4*(8), 608–620. https://doi.org/10.1038/nrmicro1462

Messier, N., & Roy, P. H. (2001). Integron integrases possess a unique additional domain necessary for activity. *Journal of Bacteriology*, *183*(22), 6699–6706. https://doi.org/10.1128/JB.183.22.6699-6706.2001

Moura, A., Henriques, I., Smalla, K., & Correia, A. (2010). Wastewater bacterial communities bring together broad-host range plasmids, integrons and a wide diversity of uncharacterized gene cassettes. *Research in Microbiology*, *161*(1), 58–66. https://doi.org/10.1016/j.resmic.2009.11.004

Moura, A., Soares, M., Pereira, C., Leitão, N., Henriques, I., & Correia, A. (2009). INTEGRALL: A database and search engine for integrons, integrases and gene cassettes. *Bioinformatics (Oxford, England)*, *25*(8), 1096–1098. https://doi.org/10.1093/bioinformatics/btp105

Nemergut, D. R., Martin, A. P., & Schmidt, S. K. (2004). Integron diversity in heavy-metal-contaminated mine tailings and inferences about integron evolution. *Applied and Environmental Microbiology*, *70*(2), 1160–1168. https://doi.org/10.1128/AEM.70.2.1160-1168.2004

Nield, B. S., Holmes, A. J., Gillings, M. R., Recchia, G. D., Mabbutt, B. C., Nevalainen, K. M., & Stokes, H. W. (2001). Recovery of new integron classes from environmental DNA. *FEMS Microbiology Letters*, *195*(1), 59–65. https://doi.org/10.1111/j.1574-6968.2001.tb10498.x

Noguchi, H., Taniguchi, T., & Itoh, T. (2008). MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Research*, *15*(6), 387–396. https://doi.org/10.1093/dnares/dsn027

Partridge, S. R., Recchia, G. D., Scaramuzzi, C., Collis, C. M., Stokes, H. W., & Hall, R. M. Y. 2000. (2000). Definition of the attI1 site of class 1 integrons. *Microbiology*, *146*(11), 2855–2864. https://doi.org/10.1099/00221287-146-11-2855

548 Pereira, M., Österlund, T., Eriksson, K. M., Backhaus, T., Axelson-Fisk, M., & Kristiansson, E.
549 (2020). A comprehensive survey of integron-associated genes present in
550 metagenomes. *BMC Genomics*, *21*(1), 495. https://doi.org/10.1186/s12864-020-
551 06830-5
552 Pielou, E. C. (1966). The measurement of diversity in different types of biological collections.
553 *Journal of Theoretical Biology*, *13*, 131–144. https://doi.org/10.1016/0022-
554 5193(66)90013-0
555 Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web
556 server: 2018 update. *Nucleic Acids Research*, *46*(W1), W200–W204.
557 https://doi.org/10.1093/nar/gky448
558 Robinson, D. G., Lee, M.-C., & Marx, C. J. (2012). OASIS: An automated program for global
559 investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Research*,
560 *40*(22), e174. https://doi.org/10.1093/nar/gks778
561 Sandoval-Quintana, E., Lauga, B., & Cagnon, C. (2022). Environmental integrons: The dark
562 side of the integron world. *Trends in Microbiology*.
563 https://doi.org/10.1016/j.tim.2022.01.009
564 Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics (Oxford,*
565 *England)*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153
566 Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical*
567 *Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
568 Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: The
569 reference centre for bacterial insertion sequences. *Nucleic Acids Research*,
570 *34*(Database issue), D32-36. https://doi.org/10.1093/nar/gkj014
571 Solovyev, V., & Salamov, A. (2011). Automatic annotation of microbial genomes and
572 metagenomic sequences. *Metagenomics and Its Applications in Agriculture,*
573 *Biomedicine and Environmental Studies*, 61–78.
574 Sonbol, S., & Siam, R. (2021). The association of group IIB intron with integrons in hypersaline
575 environments. *Mobile DNA*, *12*(1), 8. https://doi.org/10.1186/s13100-021-00234-2
576 Souque, C., Escudero, J. A., & MacLean, R. C. (2021). Integron activity accelerates the
577 evolution of antibiotic resistance. *ELife*, *10*, e62474.
578 https://doi.org/10.7554/eLife.62474
579 Stangl, C., de Blank, S., Renkens, I., Westera, L., Verbeek, T., Valle-Inclan, J. E., González,
580 R. C., Henssen, A. G., van Roosmalen, M. J., Stam, R. W., Voest, E. E., Kloosterman,
581 W. P., van Haaften, G., & Monroe, G. R. (2020). Partner independent fusion gene
582 detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore

583 sequencing. *Nature Communications*, *11*(1), 2861. https://doi.org/10.1038/s41467-
584 020-16641-7
585 Stauffert, M., Cravo-Laureau, C., Jézéquel, R., Barantal, S., Cuny, P., Gilbert, F., Cagnon, C.,
586 Militon, C., Amouroux, D., Mahdaoui, F., Bouyssiere, B., Stora, G., Merlin, F.-X., &
587 Duran, R. (2013). Impact of oil on bacterial community structure in bioturbated
588 sediments. *PLoS ONE*, *8*(6), e65347. https://doi.org/10.1371/journal.pone.0065347
589 Stokes, H. W., & Hall, R. M. (1989). A novel family of potentially mobile DNA elements
590 encoding site-specific gene-integration functions: Integrons. *Molecular Microbiology*,
591 *3*(12), 1669–1683. https://doi.org/10.1111/j.1365-2958.1989.tb00153.x
592 Stokes, H. W., Holmes, A. J., Nield, B. S., Holley, M. P., Nevalainen, K. M., Mabbutt, B. C., &
593 Gillings, M. R. (2001). Gene cassette PCR: Sequence-independent recovery of entire
594 genes from environmental DNA. *Applied and Environmental Microbiology*, *67*(11),
595 5240–5246. https://doi.org/10.1128/AEM.67.11.5240-5246.2001
596 Tanizawa, Y., Fujisawa, T., & Nakamura, Y. (2018). DFAST: A flexible prokaryotic genome
597 annotation pipeline for faster genome publication. *Bioinformatics (Oxford, England)*,
598 *34*(6), 1037–1039. https://doi.org/10.1093/bioinformatics/btx713
599 Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., & Chandler, M. (2011). ISsaga is an
600 ensemble of web-based methods for high throughput identification and semi-automatic
601 annotation of insertion sequences in prokaryotic genomes. *Genome Biology*, *12*(3),
602 R30. https://doi.org/10.1186/gb-2011-12-3-r30
603 Veltri, D., Wight, M. M., & Crouch, J. A. (2016). SimpleSynteny: A web-based tool for
604 visualization of microsynteny across multiple species. *Nucleic Acids Research*, *44*(W1),
605 W41–W45. https://doi.org/10.1093/nar/gkw330
606 von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., & Dutilh, B. E. (2019).
607 Robust taxonomic classification of uncharted microbial sequences and bins with CAT
608 and BAT. *Genome Biology*, *20*(1), 217. https://doi.org/10.1186/s13059-019-1817-x
609 Watanabe, M., Higashioka, Y., Kojima, H., & Fukui, M. (2020). Proposal of Desulfosarcina
610 ovata subsp. Sediminis subsp. Nov., a novel toluene-degrading sulfate-reducing
611 bacterium isolated from tidal flat sediment of Tokyo Bay. *Systematic and Applied*
612 *Microbiology*, *43*(5), 126109. https://doi.org/10.1016/j.syapm.2020.126109
613 Wick, R. (2018). Porechop. *Github*. https://github.com/rrwick/Porechop
614 Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2.
615 *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0
616 Zhu, W., Lomsadze, A., & Borodovsky, M. (2010). Ab initio gene identification in metagenomic
617 sequences. *Nucleic Acids Research*, *38*(12), e132. https://doi.org/10.1093/nar/gkq275
618

Data Accessibility and Benefit-Sharing

Data Accessibility Statement

The raw sequences of the intI libraries are available in GenBank under the accession numbers: FR718193-FR718248. The InOPS contig is available in Genbank under the accession number ON260918.

Benefit-Sharing Statement

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author contributions

ESQ, BL and CC conceived and designed the study. BL and CC supervised the study. CC, LH and RD collected the samples. CC and LH identified *intIOPS*. ESQ performed the CRISPR-Cas9 experiments, analyzed the data and characterized the InOPS contig. CS, GM, and IR contributed to the CRISPR-Cas9 enrichment and sequencing. BL, CC, RD and GH provided

634 fundings. ESQ, BL and CC wrote the manuscript. All authors read and approved the
635 manuscript.

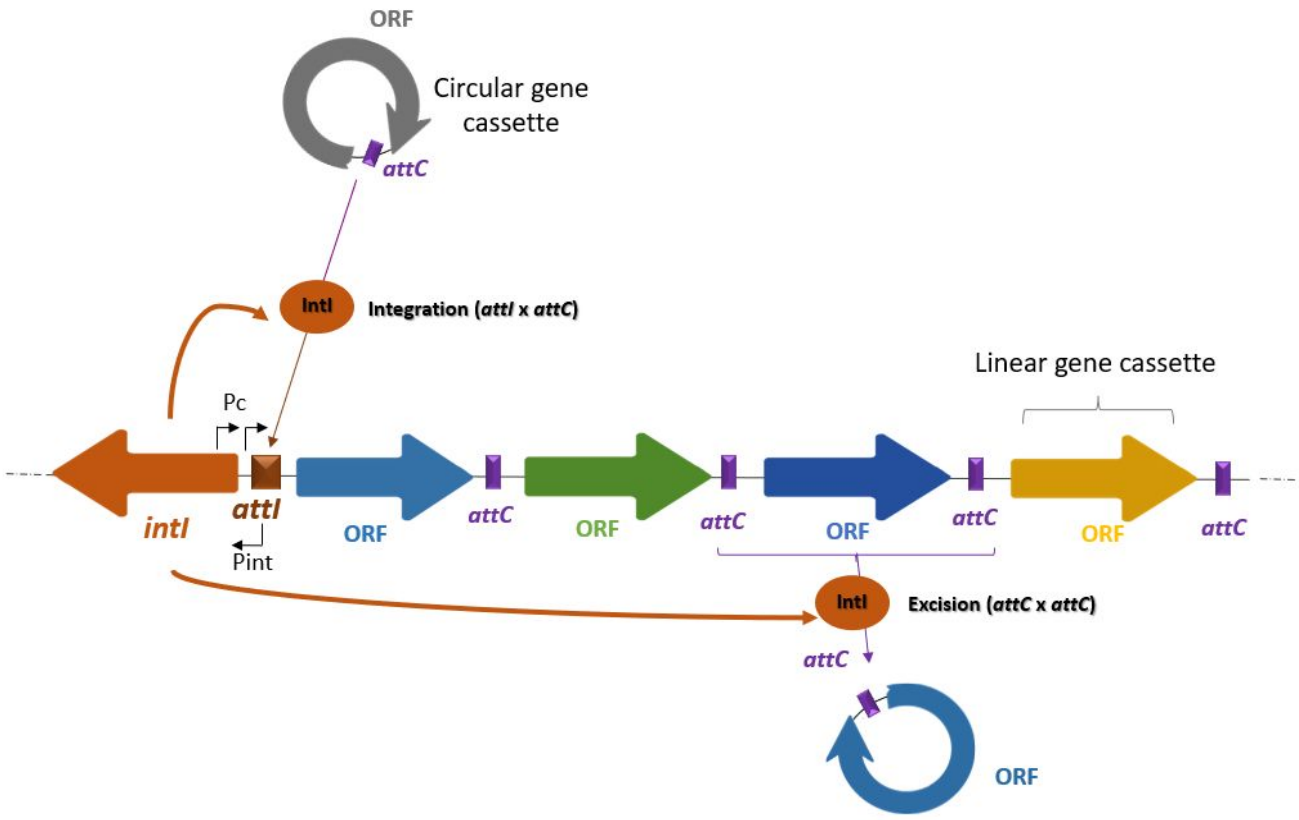
636

637

638

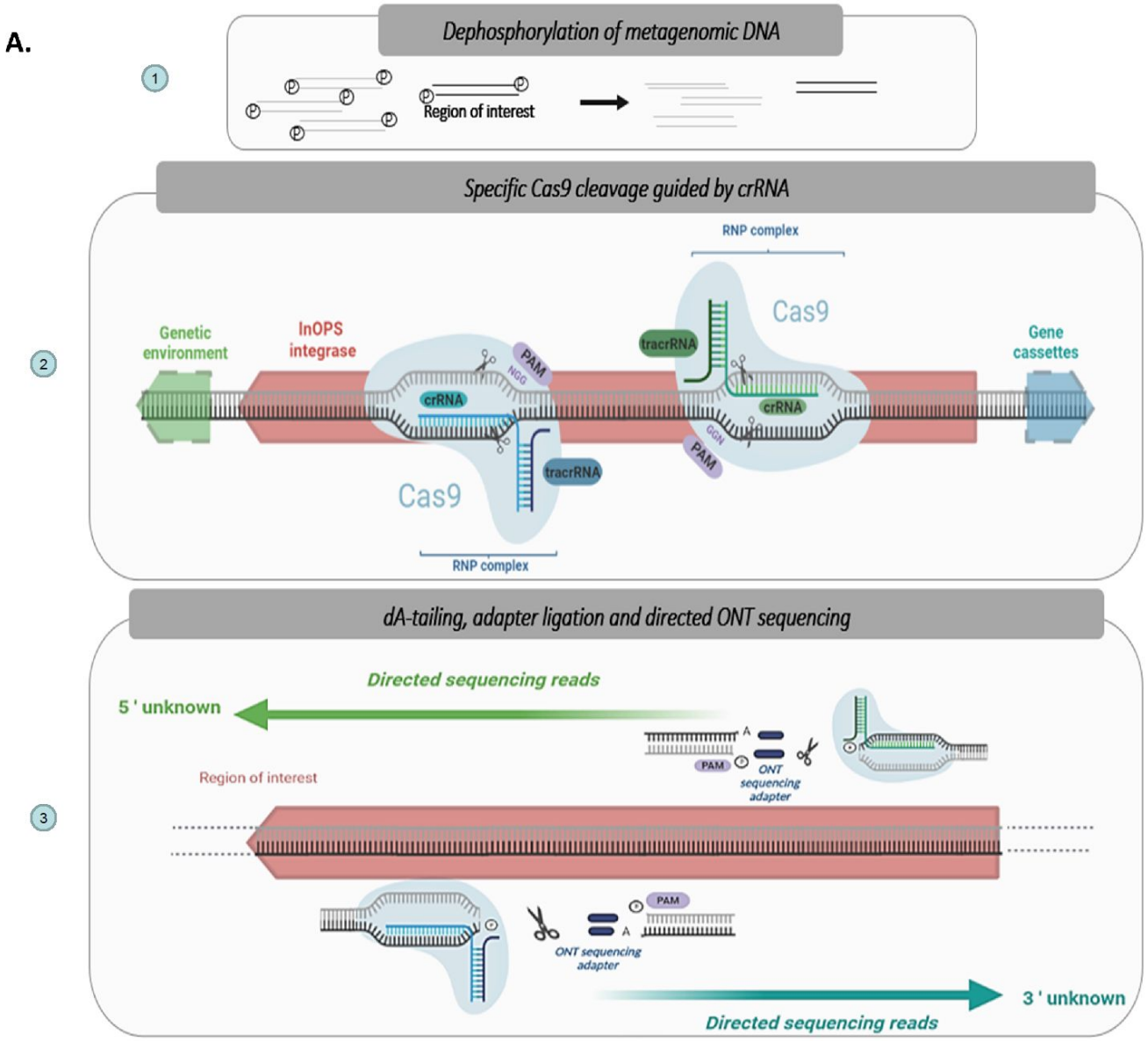
639

640 ## Figures with captions



641 **Figure 1. Integron general structure and functioning**. Integrons are formed by a functional
642 platform composed of a gene (*intI*) encoding an integrase (IntI), a promoter (Pc), and a
643 recombination site (*attI*). Usually an array of gene cassettes follows this platform. The gene
644 cassettes generally consist of an open reading frame (ORF) and a recombination site (*attC*).

The integron integrase catalyses the insertion and excision of gene cassettes by site-specific recombination. The promoters Pc (two possible locations) and PintI allow the expression of the gene cassettes and integrase, respectively.


A.
1
Dephosphorylation of metagenomic DNA
Region of interest
Specific Cas9 cleavage guided by crRNA
2
RNP complex
Genetic environment
InOPS integrase
PAM
NGG
crRNA
tracrRNA
Cas9
Gene cassettes
GGN
dA-tailing, adapter ligation and directed ONT sequencing
3
Directed sequencing reads
5 ' unknown
Region of interest
ONT sequencing adapter
3 ' unknown

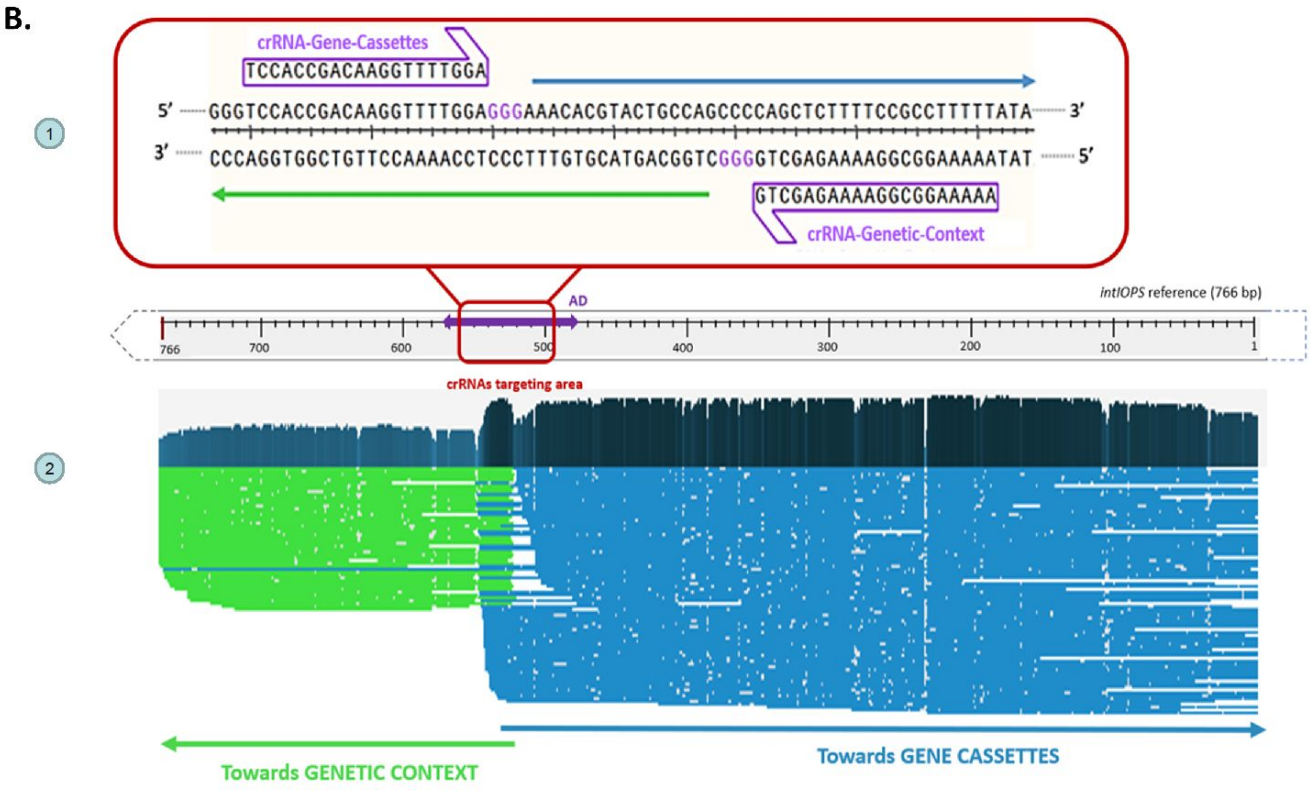
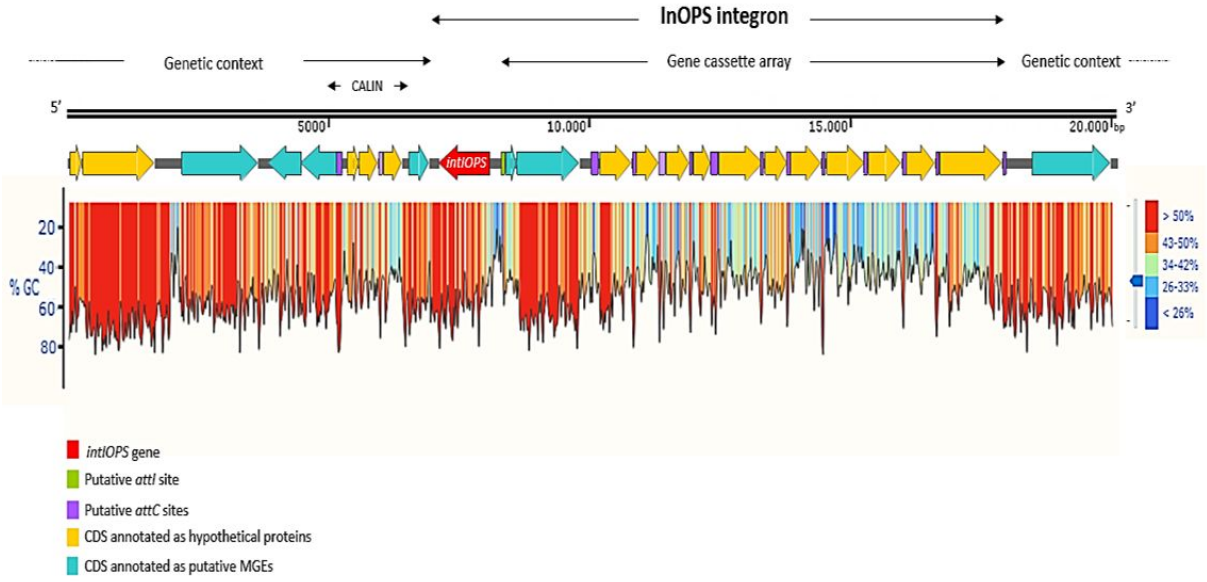

B.
1
crRNA-Gene-Cassettes
TCCACCGACAAGGTTTTGGA
5' GGGTCCACCGACAAGGTTTTGGAGGGAAACACGTACTGCCAGCCCCAGCTCTTTTCCGCCTTTTTATA 3'
3' CCCAGGTGGCTGTTCCAAAACCTCCCTTTGTGCATGACGGTCGGGGTCGAGAAAAGGCGGAAAAATAT 5'
GTCGAGAAAAGGCGGAAAAA
crRNA-Genetic-Context
AD
intIOPS reference (766 bp)
766 700 600 500 400 300 200 100 1
2
crRNAs targeting area
Towards GENETIC CONTEXT
Towards GENE CASSETTES

Figure 2. Recovery of InOPS reads by CRISPR-Cas9 enrichment and nanopore sequencing. A) General workflow of CRISPR-Cas9 enrichment. 1) Dephosphorylation of metagenomic DNA, 2) Specific Cas9 cleavages guided by the crRNAs. The two CRISPR RNA (crRNA) target a short and known sequence of the region of interest, in order to ensure the sequencing of both sides of this region in an overlapping manner. The design of crRNA requires the presence of a PAM site (5'-NGG-3') on the target sequences. The crRNA and the trans-activating CRISPR RNA (tracrRNA) link to the Cas9 nuclease constitute the ribonucleoprotein (RNP) complex. Once bound to the DNA, Cas9 produce a double-strand cleavage in the DNA in the 15-30 bp prior to the PAM site. 3) dA-tailing, adapters ligation and directed ONT sequencing. After dA-tailing, Oxford Nanopore Technology (ONT) specific sequencing adapters are ligated only to the DNA containing the PAM sequence, while the other end is blocked by Cas9 enzyme. Therefore, the sequencing is directed on only one direction spanning towards the unknown region of the targeted sequence. Merging both enrichment allows to sequence at the same time the whole region of interest. **B) Directional design of crRNA guides to target the gene *intIOPS* and read mapping:** 1) the crRNAs target a region within the additional domain (AD) of the InOPS integrase gene (*intIOPS*): positions 500 to 567 of the reference sequence (partial *intIOPS*, accession number FR718193.1). The PAM sites (5'-NGG-3') are indicated in violet. The arrows, in green and in blue, represent the direction of sequencing, towards the InOPS genetic context (5' unknown of the integron) and towards the gene cassette array (3' unknown of the integron), respectively. 2) Mapping of the recovered reads against the *intIOPS* reference sequence.



671

672 **Figure 3. InOPS contig and annotated features.** The 20 069 bp InOPS contig is presented. The
673 complete InOPS integron comprises the *intIOPS* gene (red) encoding an integron integrase
674 and a gene cassette array. The *attI* (green) and *attC* (purple if search parameter: --evalue-attC
675 1; lighter purple if search parameter: --evalue-attC 4) recombination sites are presented. A
676 CALIN (cluster of *attC* sites lacking integron-integrases) is found within the 5'-InOPS genetic
677 context. CDS encoding hypothetical proteins with no further annotation are represented in
678 yellow. The CDS with annotations related to putative MGEs are represented in blue. The GC
679 percentage along the contig is presented below the contig schema.

680

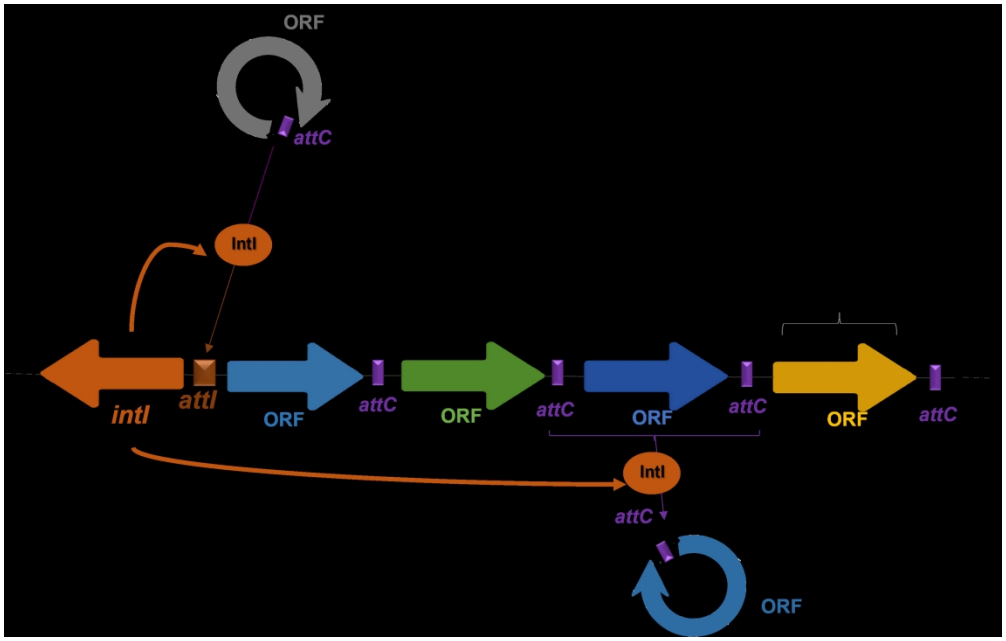
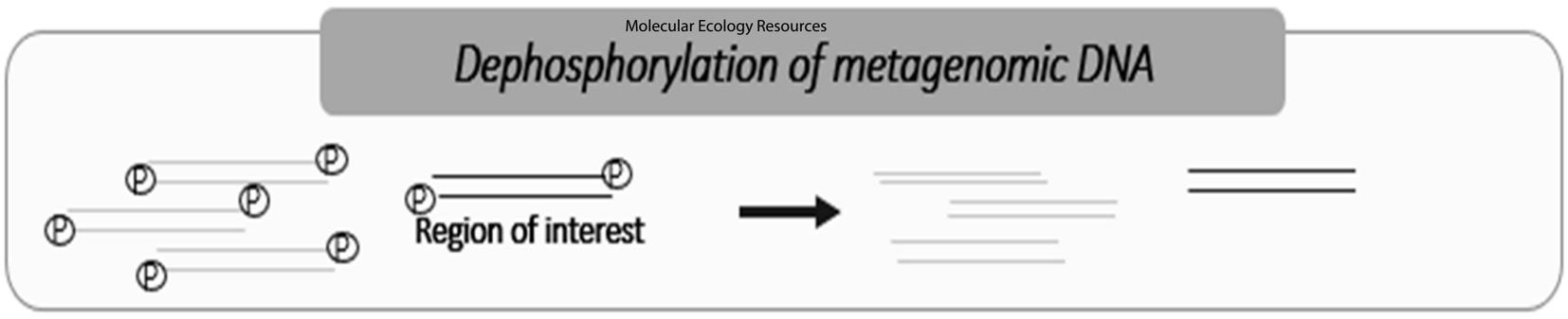


Figure 1. Integron general structure and functioning. Integrons are formed by a functional platform composed of a gene (intI) encoding an integrase (IntI), a promoter (Pc), and a recombination site (attI). Usually an array of gene cassettes follows this platform. The gene cassettes generally consist of an open reading frame (ORF) and a recombination site (attC). The integron integrase catalyses the insertion and excision of gene cassettes by site-specific recombination. The promoters Pc (two possible locations) and PintI allow the expression of the gene cassettes and integrase, respectively.

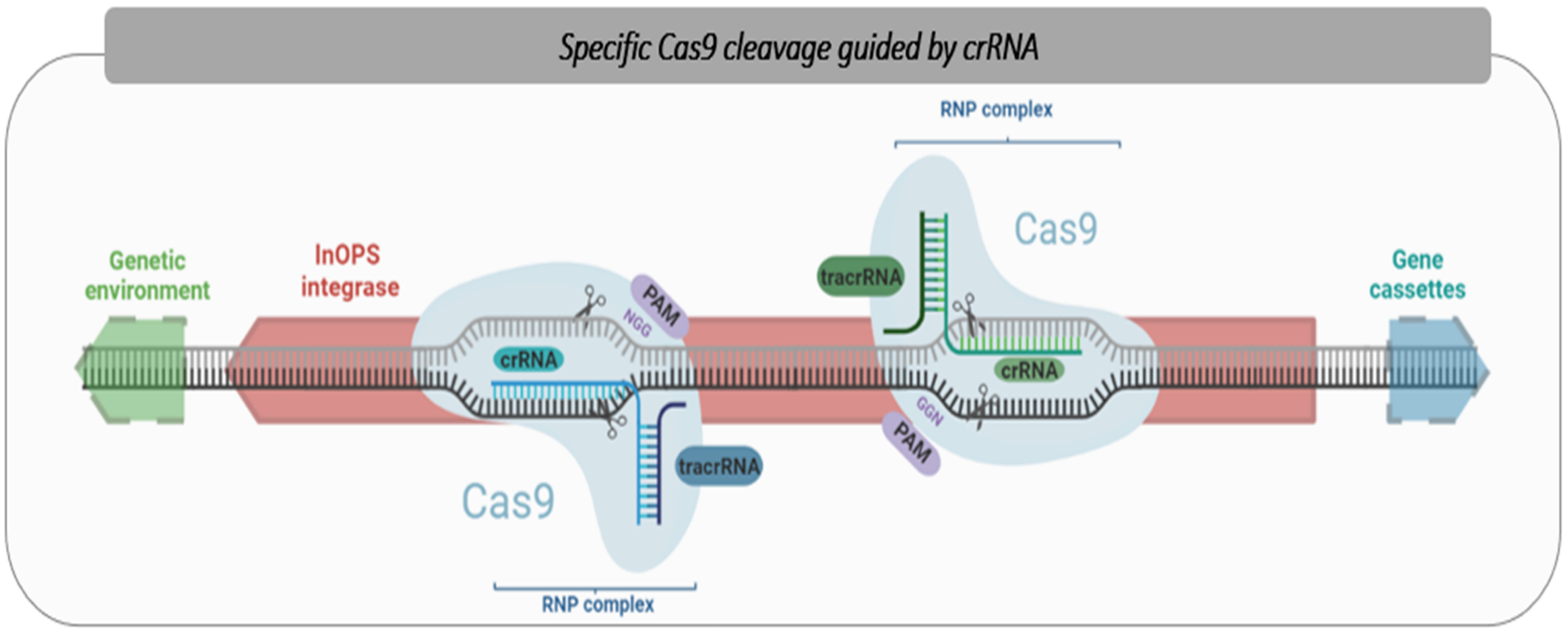
251x159mm (300 x 300 DPI)

A.

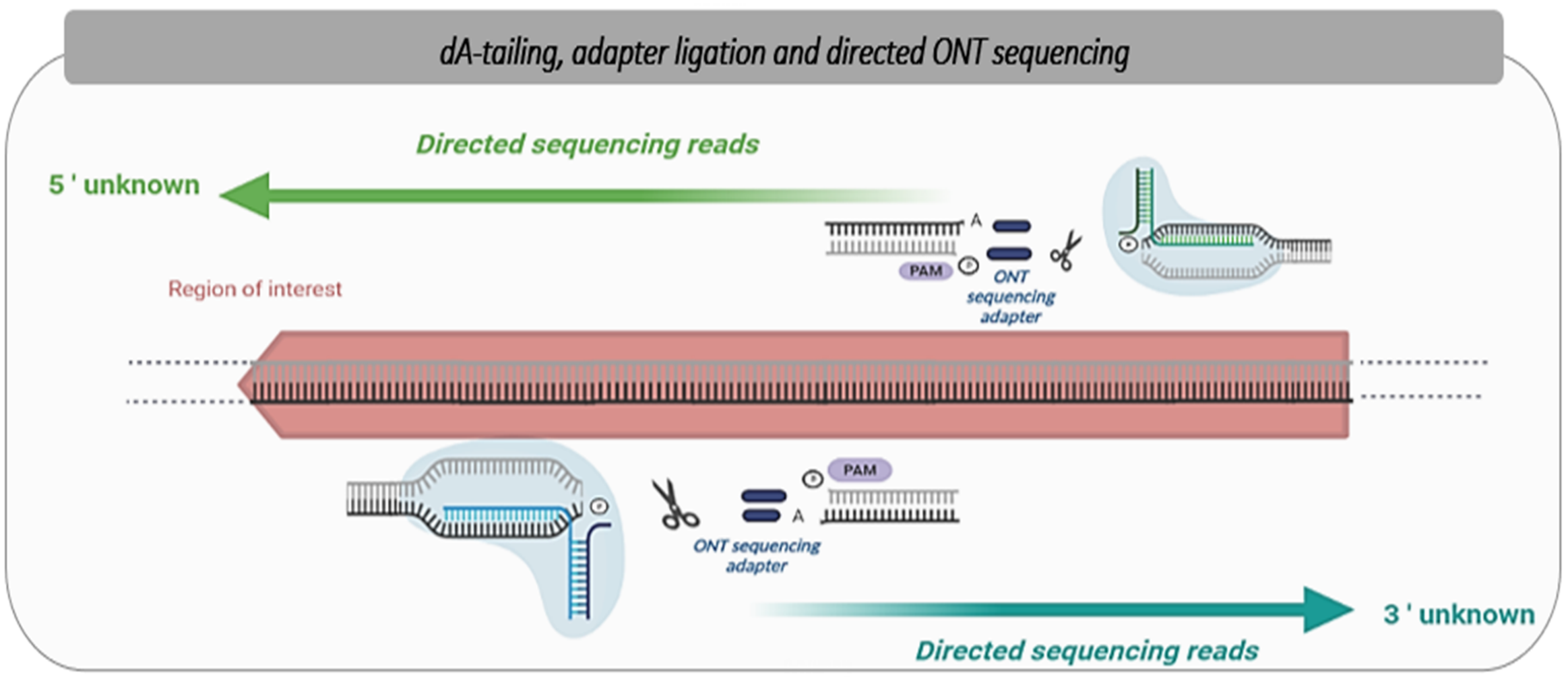
1


Dephosphorylation of metagenomic DNA
Region of interest

2


Specific Cas9 cleavage guided by crRNA
RNP complex
Genetic environment
InOPS integrase
PAM
NGG
crRNA
tracrRNA
Cas9
Cas9
tracrRNA
crRNA
GGN
PAM
Gene cassettes
RNP complex

3

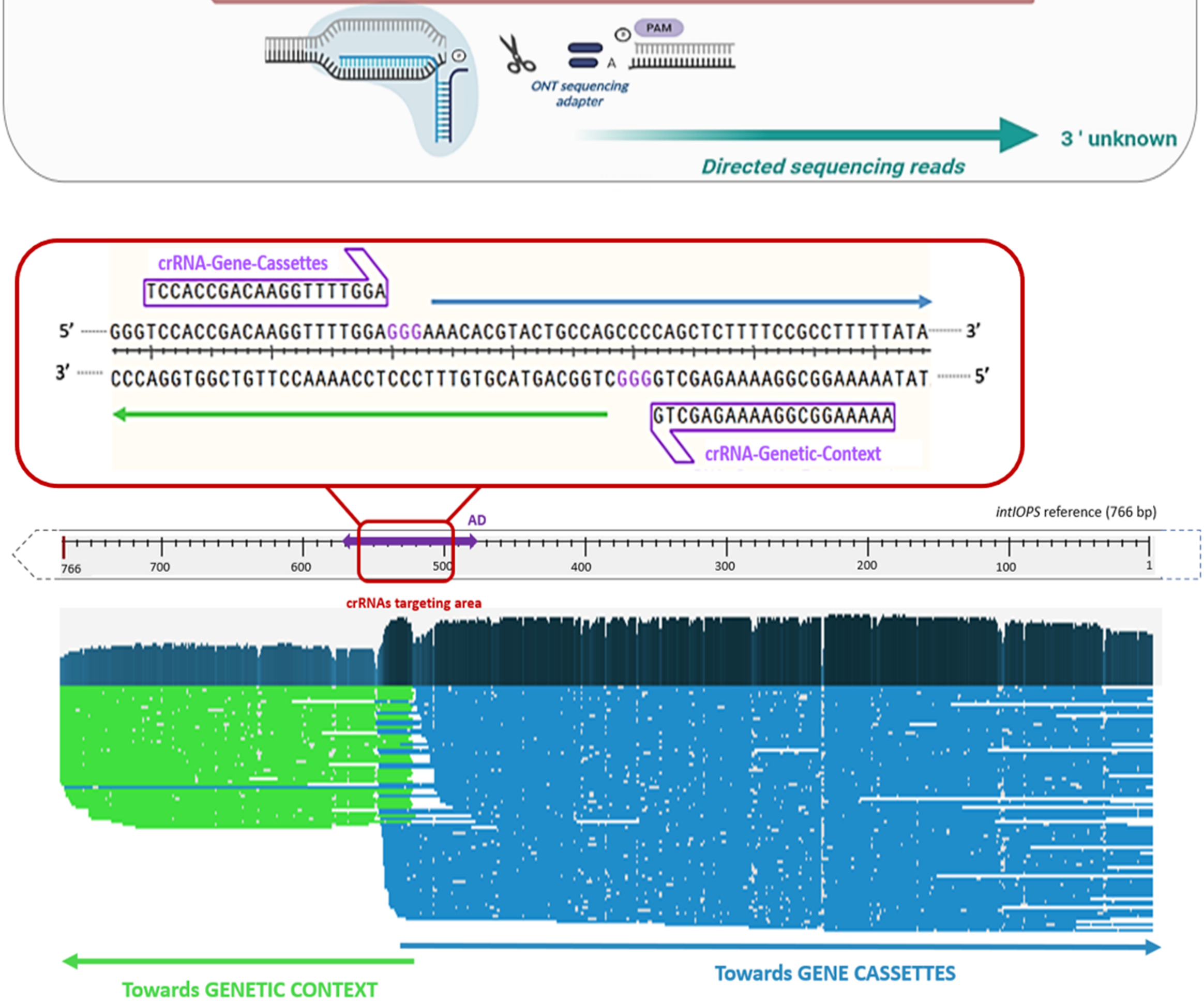

dA-tailing, adapter ligation and directed ONT sequencing
Directed sequencing reads
5 ' unknown
Region of interest
PAM
A
ONT sequencing adapter
PAM
A
ONT sequencing adapter
3 ' unknown
Directed sequencing reads

B.

1

crRNA-Gene-Cassettes
TCCACCGACAAGGTTTTGGA
5'
GGGTCCACCGACAAGGTTTTGGAGGGAAACACGTACTGCCAGCCCCAGCTCTTTTCCGCCTTTTTATA
3'
3'
CCCAGGTGGCTGTTCCAAAACCTCCCTTTGTGCATGACGGTCGGGGTCGAGAAAAGGCGGAAAAATAT
5'
GTCGAGAAAAGGCGGAAAAA
crRNA-Genetic-Context

2


AD
intIOPS reference (766 bp)
766
700
600
500
400
300
200
100
1
crRNAs targeting area
Towards GENETIC CONTEXT
Towards GENE CASSETTES

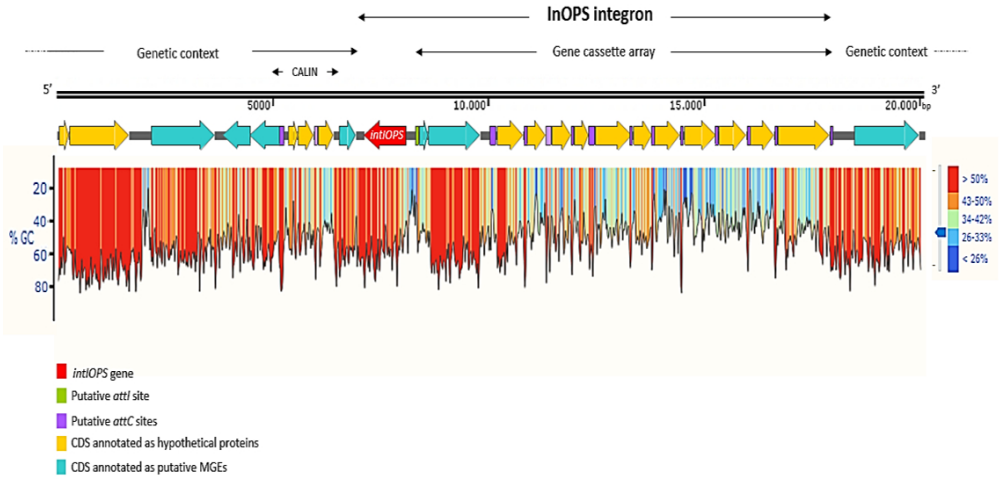


Figure 3. InOPS contig and annotated features. The 20 069 bp InOPS contig is presented. The complete InOPS integron comprises the intIOPS gene (red) encoding an integron integrase and a gene cassette array. The attI (green) and attC (purple if search parameter: --evalue-attC 1; lighter purple if search parameter: --evalue-attC 4) recombination sites are presented. A CALIN (cluster of attC sites lacking integron-integrases) is found within the 5'-InOPS genetic context. CDS encoding hypothetical proteins with no further annotation are represented in yellow. The CDS with annotations related to putative MGEs are represented in blue. The GC percentage along the contig is presented below the contig schema.

300x151mm (96 x 96 DPI)