MICROBIAL GENOMICS

METHOD

Ghaly *et al.*, *Microbial Genomics* 2022;8:000788

DOI 10.1099/mgen.0.000788

# Methods for the targeted sequencing and analysis of integrons and their gene cassettes from complex microbial communities

Timothy M. Ghaly[1,*], Anahit Penesyan[1,2], Alexander Pritchard[3], Qin Qi[1], Vaheesan Rajabal[1,2], Sasha G. Tetu[1,2] and Michael R. Gillings[1,2]

## Abstract

Integrons are microbial genetic elements that can integrate mobile gene cassettes. They are mostly known for spreading antibiotic resistance cassettes among human pathogens. However, beyond clinical settings, gene cassettes encode an extraordinarily diverse range of functions important for bacterial adaptation. The recovery and sequencing of cassettes has promising applications, including: surveillance of clinically important genes, particularly antibiotic resistance determinants; investigating the functional diversity of integron-carrying bacteria; and novel enzyme discovery. Although gene cassettes can be directly recovered using PCR, there are no standardised methods for their amplification and, importantly, for validating sequences as genuine integron gene cassettes. Here, we present reproducible methods for the amplification, sequence processing, and validation of gene cassette amplicons from complex communities. We describe two different PCR assays that either amplify cassettes together with integron integrases, or gene cassettes together within cassette arrays. We compare the performance of Nanopore and Illumina sequencing, and present bioinformatic pipelines that filter sequences to ensure that they represent amplicons from genuine integrons. Using a diverse set of environmental DNAs, we show that our approach can consistently recover thousands of unique cassettes per sample and up to hundreds of different integron integrases. Recovered cassettes confer a wide range of functions, including antibiotic resistance, with as many as 300 resistance cassettes found in a single sample. In particular, we show that class one integrons are collecting and concentrating resistance genes out of the broader diversity of cassette functions. The methods described here can be applied to any environmental or clinical microbiome sample.

## DATA SUMMARY

The authors confirm that all supporting data and protocols have been provided within the article or through supplementary data files. The code used for filtering sequences to ensure that they represent amplicons from genuine integrons are available at https://github.com/timghaly/integron-filtering. The code used to predict the taxonomic sources of gene cassette recombination sites (*attC*s) is available at https://github.com/timghaly/attC-taxa. All raw sequence data are available from the NCBI SRA database under the BioSample accessions SAMN21354384 to SAMN21354431. All BioSamples are linked to the NCBI BioProject PRJNA761546.

## INTRODUCTION

Integrons are microbial genetic elements that can capture, mobilise, and rearrange gene cassettes [1, 2]. They are mostly known for spreading a diverse repertoire of gene cassettes that collectively confer resistance to almost all classes of antibiotics [3]. Beyond clinical settings, however, integrons play a crucial role in bacterial evolution by rapidly generating genomic diversity [4, 5]. Functional integrons are characterised by their flagship gene, the integron integrase (*intI*), which encodes a site-specific tyrosine recombinase (IntI). IntI mediates the insertion of gene cassettes at the integron recombination site (*attI*), which acts as the insertion site of captured gene

Received 19 September 2021; Accepted 27 January 2022; Published 17 March 2022

**Author affiliations:** [1]School of Natural Sciences, Macquarie University, New South Wales 2109, Australia; [2]ARC Centre of Excellence in Synthetic Biology, Macquarie University, New South Wales 2109, Australia; [3]Division of Food Sciences, University of Nottingham, Loughborough LE12 5RD, Australia.

***Correspondence:** Timothy M. Ghaly, timothy.ghaly@mq.edu.au

**Keywords:** Gene cassette PCR; class 1 integrons; sedentary chromosomal integrons; antibiotic resistance.

**Abbreviations:** CM, covariance model; COG, Clusters of Orthlogous Groups; HMM, hidden Markov model; ONT, Oxford Nanopore Technologies; ORF, open reading frame; SCI, sedentary chromosomal integron.

**Data statement:** One supplementary table and five supplementary figures are available with the online version of this article.

**Impact Statement**

Integrons are microbial genetic elements that can rapidly generate genetic diversity by inserting and rearranging modular, mobilisable genes known as gene cassettes. Integrons and their gene cassettes are extensively studied around the world, with a particular focus on the diverse repertoire of antimicrobial resistance genes that they disseminate globally. Currently, a PCR-based approach represents the most efficient method of studying integrons within complex microbial communities. However, there are no standardised methods for this approach, and in particular, there lacks any robust methods for validating sequences as genuine integrons. This has serious implications for the reliability of any biological conclusions drawn from such studies. Here, we present reproducible experimental and computational methods for the PCR amplification, sequence processing, and the validation of gene cassette sequences from complex communities. The methods presented here will aid future studies of integrons and their associated gene cassettes, and will help foster reliable and standardised results.

cassettes [6]. Gene cassettes, prior to their insertion, are circular molecules, which possess a cassette recombination site (*attC*). Their insertion involves IntI-mediated recombination between the *attI* site of the integron and the *attC* site of the cassette [7–11]. Multiple cassettes can be inserted to form a linear cassette array, which can vary in size from zero to hundreds [12, 13]. IntI activity is induced by DNA damage, often triggered by environmental stress [14, 15]. Integrons can therefore provide genomic diversity at precisely the moment when it is needed the most, thus facilitating 'adaptation on demand' [16].

Recovery and sequence analysis of integron gene cassettes can serve several purposes. First, screening gene cassettes can provide a direct method for surveillance of resistance genes that are prevalent in an environment of interest. It has also been proposed that surveying gene cassettes can help detect novel functions that might be harmful to human health, such as increased pathogenicity or resistance to novel antibiotics [17]. In particular, class 1 integrons, due to their mobility, abundance, and distribution [18, 19], are primed to play a crucial role in dissemination of these genes. Finally, exploring gene cassettes provides a window into the functional diversity of the bacterial pangenome. Gene cassettes have been found to be extraordinarily abundant and diverse in every environment surveyed [20–26]. Further, many cassettes with known functions act as single-gene/single-trait entities [17, 27]. As such, they need minimal integration into metabolic networks and can likely function in a relatively wide range of genomic contexts. These traits make them highly valuable commodities for synthetic biology and biotechnological applications, particularly for the discovery of diverse enzymatic activities [17].

Currently, gene cassettes can be recovered from genome sequencing of cultured isolates, whole metagenomic sequencing, or amplicon sequencing of *attC*-associated genes. Since most bacteria currently remain unculturable, cassettes identified from isolate genomes inevitably reflect only a small proportion of all gene cassettes, exacerbated by the fact that different strains of the same species can vary widely in cassette content [4]. Whole metagenomic sequencing, although potentially a less biased approach, can be challenging, as gene cassettes often exist at very low abundances and can contain repeat sequences. A targeted amplicon sequencing approach, however, can overcome these issues and could provide the most efficient method for recovering diverse gene cassettes from complex microbial communities [28].

Cassette-targeted amplicon sequencing has been used previously, with varying returns in gene cassette recovery [20–26, 28]. As sequencing technologies have improved, the ability to capture a greater diversity of gene cassettes has also increased [20, 26]. However, such studies lack standardised methods for amplifying and, importantly, validating amplicon sequences as part of genuine cassettes arrays. Due to the degenerate nature of integron-targeting PCR primers, off-target amplification poses a serious threat to gaining trustworthy and biologically meaningful data. As such, we present standardised and reproducible methods for amplifying, sequencing, and stringent bioinformatic filtering of genuine gene cassettes from mixed microbial communities.

We applied two different PCR assays using DNA isolated from diverse environmental samples with the aim of recovering integron integrases and gene cassettes. PCR products were sequenced with both long-read Oxford Nanopore (ONT) and short-read Illumina MiSeq sequencing technologies. Importantly, we present bioinformatic pipelines that filter sequences for complete *attC* sites or *intI* genes to obtain genuine, high-quality integron sequence data. We show that after filtering, we can consistently recover thousands of gene cassettes from a single sample. We find that recovered putative genes encode a diverse suite of functional traits, including antibiotic resistance.

## METHODS

### Sample collection and DNA extraction

Duplicate samples were collected from six different sites (three terrestrial and three aquatic environments). Terrestrial sites consisted of urban parkland soil from Macquarie University (Sydney, New South Wales, Australia) [20], hot desert soil from Sturt National Park (North-western New South Wales) [29, 30], and Antarctic soil from Herring Island [20]. The aquatic sites

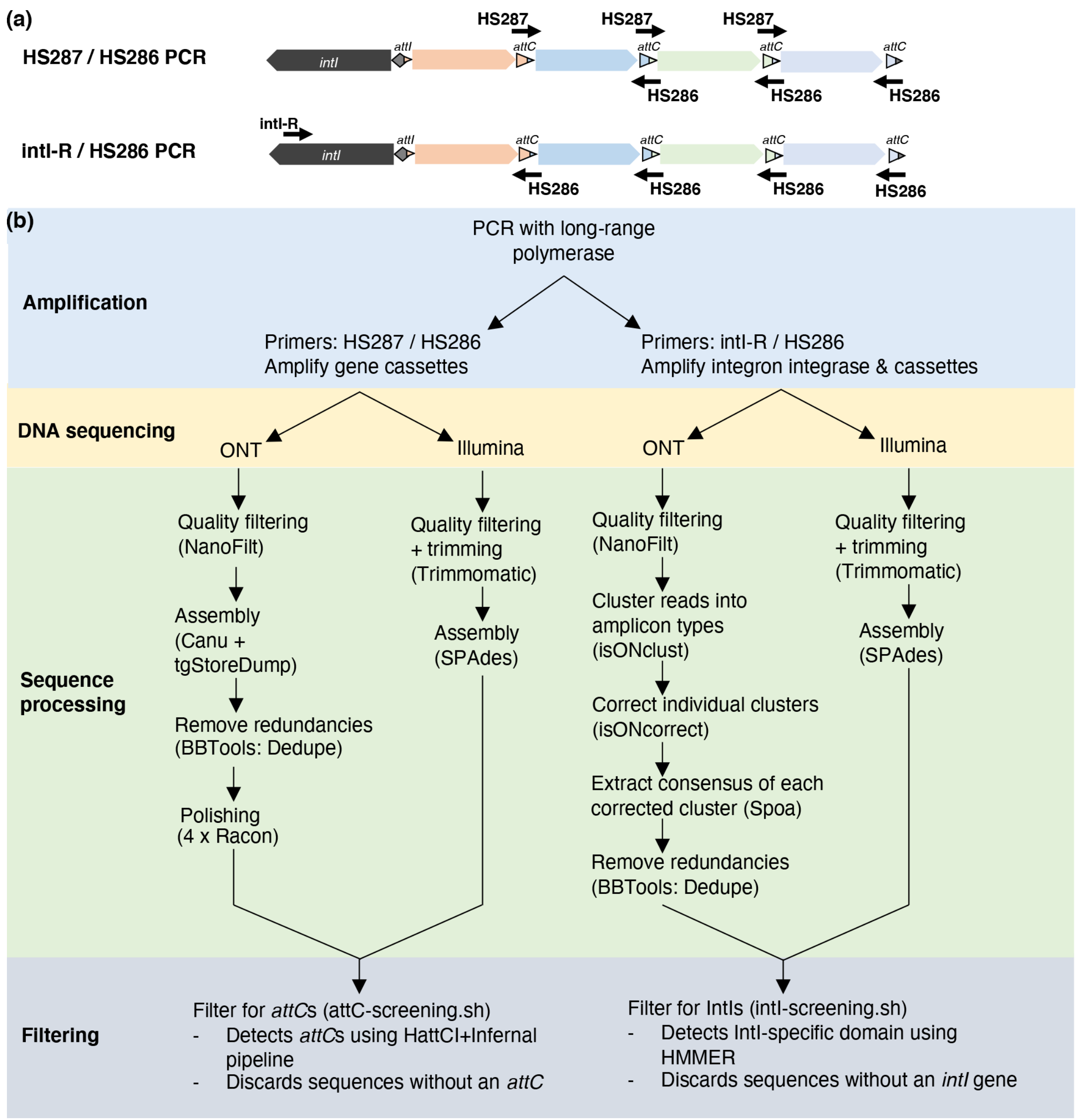
(a)
HS287 / HS286 PCR
HS287
HS287
HS287
intI
attI
attC
attC
attC
attC
HS286
HS286
HS286
intI-R / HS286 PCR
intI-R
intI
attI
attC
attC
attC
attC
HS286
HS286
HS286
HS286
(b)
PCR with long-range polymerase
Amplification
Primers: HS287 / HS286
Amplify gene cassettes
Primers: intI-R / HS286
Amplify integron integrase & cassettes
DNA sequencing
ONT
Illumina
ONT
Illumina
Sequence processing
Quality filtering (NanoFilt)
Assembly (Canu + tgStoreDump)
Remove redundancies (BBTools: Dedupe)
Polishing (4 x Racon)
Quality filtering + trimming (Trimmomatic)
Assembly (SPAdes)
Quality filtering (NanoFilt)
Cluster reads into amplicon types (isONclust)
Correct individual clusters (isONcorrect)
Extract consensus of each corrected cluster (Spoa)
Remove redundancies (BBTools: Dedupe)
Quality filtering + trimming (Trimmomatic)
Assembly (SPAdes)
Filtering
Filter for attCs (attC-screening.sh)
- Detects attCs using HattCI+Infernal pipeline
- Discards sequences without an attC
Filter for IntIs (intI-screening.sh)
- Detects IntI-specific domain using HMMER
- Discards sequences without an intI gene

**Fig. 1.** Experimental and bioinformatic workflow for gene cassette amplicon sequencing. (a) Components of integrons amplified by the two PCR assays. The primer set HS287/HS286 targets cassettes that lie between two *attC* sites. Potentially any gene cassette(s) can be amplified by this primer set. The primer set intI-R / HS286 targets diverse integron integrases (*intI*) and cassette recombination sites (*attC*). The resulting amplicons include ~800 bp of *intI* and at least the first cassette(s) of an array. (b) The bioinformatic steps and software (in parentheses) used to process and filter amplicon data. Methods are shown for both primer sets sequenced with either Oxford Nanopore (ONT) or Illumina technologies.

consisted of river sediment (Lane Cove River, New South Wales), freshwater biofilms (Mars Creek, New South Wales) [31], and estuarine sediment (Paramatta River Estuary, New South Wales). From each of the 12 samples, DNA was extracted from 0.3 g of material using a standard bead-beating protocol [32]. Each resulting DNA sample was used as the template for two different PCR assays, described below, and all were subsequently sequenced using long-read Oxford Nanopore (ONT) and short-read MiSeq (Illumina) technologies (Fig. 1).

## PCR amplification and DNA sequencing

For each sample, we conducted two different PCR assays (Fig. 1a). The first used the primers HS287 and HS286 [28], which target *attC* recombination sites in opposing directions to amplify intervening gene cassettes. The second primer set, intI-R / HS286, amplifies approximately 800 bp of the integron integrase gene as well as downstream gene cassettes. The primer intI-R (5'- GCG

AAC GAR TGB CGV AGV GTG TG –3') was designed to target diverse integron integrases and was based on an alignment of 174 complete *intI* sequences containing a functional catalytic site, as compiled by Cambray *et al.* [14]. Importantly, the last 6 bp of the 3′ end of intI-R exactly matches 75% of aligned *intI* sequences. For amplification, we used Phusion Hot Start II DNA Polymerase (ThermoFisher Scientific, Waltham, MA, USA), which is a long-range DNA polymerase, chosen to facilitate the amplification of large segments of integron cassette arrays, known to reach more than 100 kilobases in length [12]. The PCRs were carried out in 50 µL volumes containing a final concentration of 1 x GC Phusion Buffer, 0.2 mM dNTPs, 0.5 µM of each primer, 3% DMSO and 2 U of Phusion DNA polymerase. All PCRs were performed using GeneReleaser (Bioventures, Murfreesboro, TN, USA) as previously described [33]. Triplicate PCRs were performed for each sample to increase the chances of capturing rare gene cassettes that might otherwise escape amplification due to the stochastic nature of PCR.

For the HS287/HS286 primer set, the following thermal cycling programme was used: 98 °C for 3 min for one cycle; 98 °C for 10s, 60 °C for 30 s, 72 °C for 3 min 30 s for 35 cycles; and a final extension step at 72 °C for 10 min. For the intI-R / HS286 primer set, the following thermal cycling programme was used: 98 °C for 3 min for one cycle; 98 °C for 10 s, 65 °C for 30 s, 72 °C for 3 min 30 s for 35 cycles; and a final extension step at 72 °C for 10 min. PCR efficiency was assessed using 2% agarose gel electrophoresis. Triplicate PCRs were pooled and then purified with AMPure XP beads (Beckman Coulter, Danvers, MA, USA) using a 1.8:1 beads-to-sample ratio as per the manufacturer's protocol.

For ONT sequencing, the 24 PCR products (representing the 12 samples amplified with each primer set) were multiplexed in a single DNA library using the ONT Ligation Sequencing Kit (SQK-LSK109) and the ONT Native Barcoding Expansion Kits (EXP-NBD104 and EXP-NBD114) according to the manufacturer's protocol. The DNA library was sequenced using a MinION MK 1B sequencing device on an R10.3 flow cell. Sequencing was allowed to run for 24 h. Basecalling was carried out with Guppy v.4.3.4 [34] with default parameters using the high accuracy (HAC) basecalling model.

For short-read sequencing, the 24 PCR products underwent an Illumina DNA shotgun library preparation using the Nextera XT protocol and then sequenced with MiSeq 300 bp paired-end sequencing on a single lane. Illumina sequencing and library preparation were carried out at the Australian Genome Research Facility (Melbourne, Australia).

### Sequence processing and *attC* filtering: HS287 / HS286 PCRs

To compare the performance of short- and long-read sequencing technologies, we sequenced HS287/HS286 amplicons on both Nanopore (ONT) and Illumina platforms. The respective workflows and software used for sequence processing and filtering are summarised in Fig. 1(b).

For ONT sequences of the HS287/HS286 PCRs, we first filtered reads based on average quality (q) scores. We removed any reads with an average q score below 10 using NanoFilt v2.8 [35] [parameters: -q 10]. Although each read spans the length of an entire amplicon, many amplicons represent overlapping subsections of larger potential templates. Thus, an assembly of these initial reads into larger cassette arrays was carried out using Canu v2.0 [36] [parameters: genomeSize=5m minReadLength=250 minOverlapLength=200 corMinCoverage=0 corOutCoverage=20000 corMhapSensitivity=high maxInputCoverage=20000 batMemory=125 redMemory=32 oeaMemory=32 batThreads=24 purgeOverlaps=aggressive]. Assembled contigs and unassembled reads were then extracted together using the tgStoreDump script within Canu [parameters: -consensus -fasta]. Any redundancies were removed using dedupe.sh, available from the BBTools package v35 [37] with default parameters. Consensus sequences were then corrected with four rounds of polishing using Racon v1.4.20 [38]. Each round of Racon polishing involved read mapping with minimap2 v2.20-r1061 [39] [parameters: -x map-ont -t 24] and error correction with Racon [parameters: -m 8 -x 6 -g -8 -w 500 -t 24].

For Illumina sequence data of the HS287/HS286 PCRs, paired-end reads first underwent quality trimming and adapter clipping using Trimmomatic v0.38 [40] [parameters: -phred33 ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30] where 'adapters.fa' is a fasta-formatted file containing all commonly used Illumina adapter sequences. If only one end of paired reads had acceptable quality, it was used as a single read during assembly. Surviving paired-end reads and single reads were assembled together using SPAdes v3.14.1 [41–43] [parameters -k 21,33,55,77,99,127 --only-assembler --careful].

The resulting ONT and Illumina sequences were both filtered based on the presence of internal cassette recombination sites (*attC*s). Given the degenerate nature of the primers, off-target amplicons might constitute a significant portion of the reads. Filtering for sequences that have internal *attC* sites is thus an essential step when analysing cassette amplicon data. It should be noted that filtering for *attC*s in this way may discard some genuine amplicons which consist of single gene cassettes, since they do not possess a complete *attC* site (Fig. 1a). Nevertheless, we consider that for obtaining meaningful ecological data, the removal of potential false positives is more important than the loss of some true positives. To filter for *attC* sites, we used an in-house script, attC-screening.sh (available: https://github.com/timghaly/integron-filtering), with default parameters. The script uses the HattCI [44]+Infernal [45] pipeline that has been previously described [46, 47]. In short, attC-screening.sh searches for the sequence and

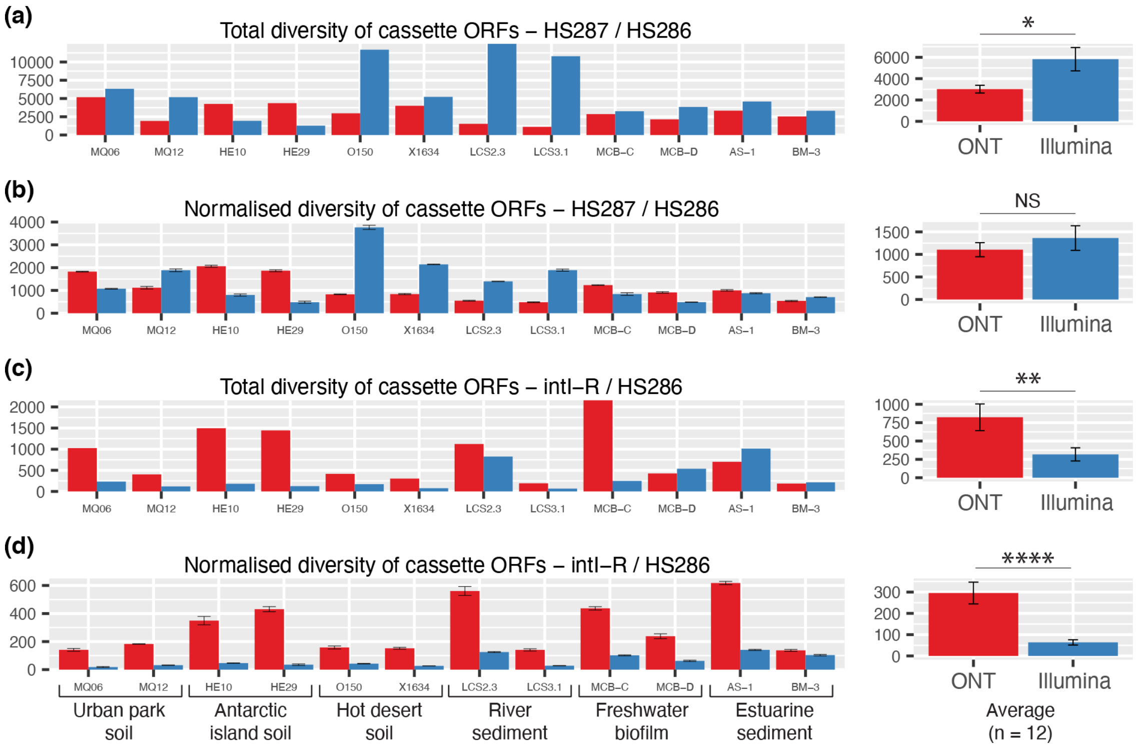

**Fig. 2.** Diversity of recovered gene cassette ORFs. Redundancy was removed using a 100% amino acid identity of translated protein sequences. (a) Total non-redundant cassette ORFs amplified using the primers HS287/HS286. (b) Cassette ORF diversity was normalised for sequencing depth, based on averages (±1 S.E) of triplicate 50-megabase subsamples of raw sequence reads. Total (c) and normalised (d) cassette ORF diversity are shown for the intI-R / HS286 primer set. Average (±1 S.E) diversity for each analysis are shown on the right-hand side of each panel. The degree of statistical significance is shown by asterisks as determined by two-sample T-tests or Wilcoxon rank sum tests (depending on the normality of the data). NS: *P*>0.05, *: *P*<0.05, **: *P*<0.01, ***: *P*<0.001, ****: *P*<0.0001.

secondary structures conserved among *attC*s and retains any input sequence that has at least one *attC* site. The script can be used on fasta-formatted data generated from any sequencing technology.

## intI-R / HS286 PCRs: sequence processing and IntI filtering

All intI-R / HS286 PCRs were also sequenced on both ONT and Illumina platforms (Fig. 1b). For Nanopore sequencing, basecalled reads were first quality filtered using NanoFilt v2.8 [35] [parameters: -q 10]. Reads representing concatemers and chimaeras were removed using yacrd v0.6.2 [48] with default parameters for ONT data. Since all amplicons should be anchored on one end to the *intI* gene, an assembly would not be suitable. Instead, we clustered reads into amplicon 'types' using isONclust v0.0.6.1 [49]. Each cluster was then individually corrected using isONcorrect v0.0.8 [50] with default parameters. Unlike error-correction of genomic data, isONcorrect takes into account uneven coverage within the same read as well as structural variation among similar reads from different clusters (e.g. reads that represent true biological rearrangements of the same gene cassettes). From each corrected cluster, a consensus sequence was then generated using spoa v4.0.7 [38] [parameter: -r 0]. Any redundancies were removed using the BBTools v35 [37] script dedupe.sh with default parameters.

Illumina sequences were processed in the same manner as described above for the HS287/HS286 data. This involved quality trimming and adapter clipping using Trimmomatic v0.38 [40], followed by an assembly of the reads using SPAdes v3.14.1 [41–43].

The resulting ONT and Illumina sequences were both filtered based on the presence of IntI protein sequences. To detect sequences that encoded IntI, we used an in-house script, intI-screening.sh (available: https://github.com/timghaly/integron-filtering), with default parameters. The script uses a profile hidden Markov model (HMM) provided by Cury *et al.* [13] to detect the additional domain that is unique to integron integrases (I2 α-helix), separating them from other tyrosine recombinases [51, 52]. The intI-screening.sh pipeline first uses Prodigal [53] to predict all encoded protein sequences, and then screens them for the IntI-specific domain using hmmsearch from the HMMER v3 software package [54]. Any sequences that do not contain a recognisable integron integrase are discarded. Similarly, intI-screening.sh can be used on fasta-formatted data generated from any sequencing technology.

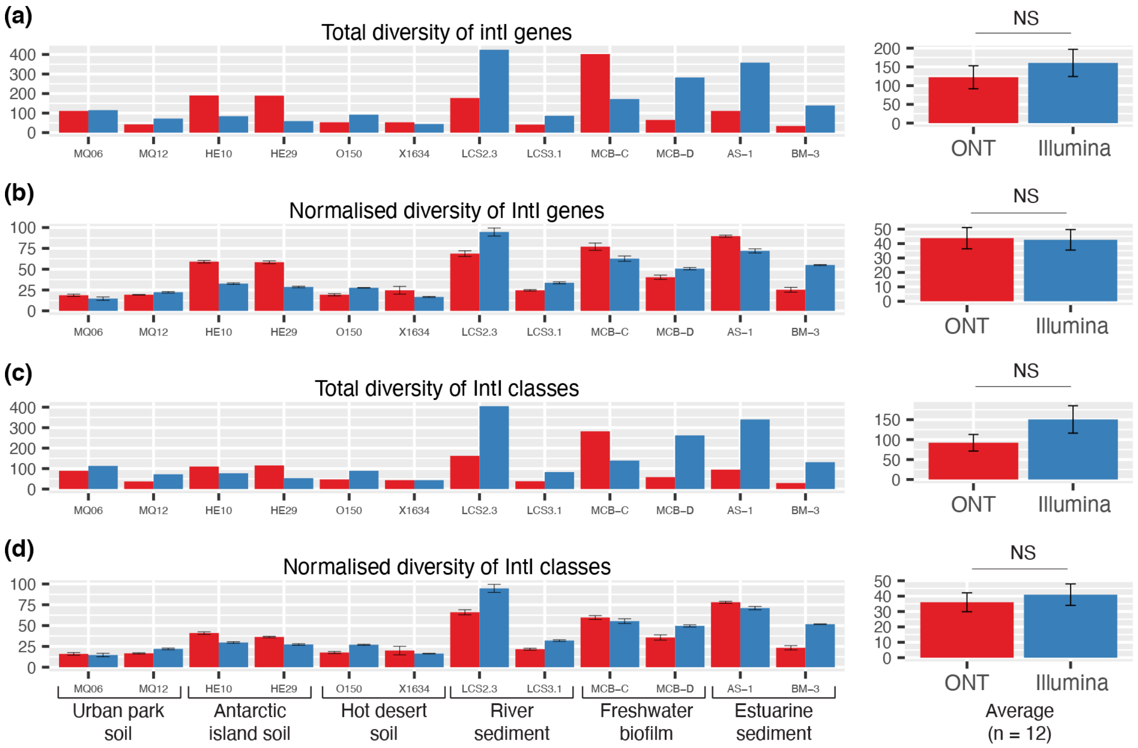

**Fig. 3.** Diversity of integron integrases recovered by the intI-R / HS286 primer set. (a) Total non-redundant (100% amino acid identity) integron integrases (IntIs) recovered. (b) IntI diversity was normalised for sequencing, based on averages (±1 S.E) of triplicate 50-megabase subsamples of raw sequence reads. Total (c) and normalised (d) diversity of IntI classes (using a 94% amino acid clustering threshold) are shown. Average (±1 S.E) diversity for each analysis are shown on the right-hand side of each panel. Differences between Nanopore (ONT) and Illumina MiSeq technologies were not significant (NS) as determined by Wilcoxon rank sum tests.

## Protein prediction and functional classification of gene cassettes

Cassette open reading frames (ORFs) and their translated protein sequences were predicted using Prodigal v2.6.3 [53] in metagenomic mode [parameters: -p meta].

To assess the broad-scale functional diversity of gene cassettes, we used the Clusters of Orthologs Groups (COGs) database [55]. COG functions were assigned to cassette-encoded protein sequences using eggNOG-mapper v2.0.1b [56, 57] executed in DIAMOND [58] mode with default parameters. To detect cassette-encoded antimicrobial resistance genes (ARGs), we used ABRicate v0.8 [59] to search against the Comprehensive Antibiotic Resistance Database (CARD) v3.1.1 [60] [Downloaded: 21 April 2021].

## Taxonomic classification of *attC* sites

The gene cassettes of sedentary chromosomal integrons (SCIs) generally possess highly similar *attC* sites, and this conservation spans the SCIs of different bacteria within the same taxon [4, 61, 62]. We have recently modelled the conserved sequence and structure of *attC* sites from the chromosomal integrons of 11 bacterial taxa [46]. These included six Gammaproteobacterial orders (Alteromonadales, Methylococcales, Oceanospirillales, Pseudomonadales, Vibrionales, Xanthomonadales) and an additional five phyla (Acidobacteria, Cyanobacteria, Deltaproteobacteria, Planctomycetes, Spirochaetes). A covariance model (CM) was generated separately for each taxon, and this can be used to correctly identify the source taxon of *attC* sites with high specificity (98–100%) [46].

Here, we used an in-house script, attC-taxa.sh (available: https://github.com/timghaly/attC-taxa), with default parameters to detect any *attC* sites that have originated in the SCIs of one of the 11 taxa. The attC-taxa.sh pipeline incorporates all 11 CMs and uses cmsearch [parameters: --notrunc --max] from the Infernal software package [45] to classify *attC*s. It is important to note that each taxon-specific model exhibits different sensitivities in detecting true positives and thus the relative proportion of different taxa cannot be compared within the same sample. However, the relative proportion of the same taxon can be compared between different samples.

**Table 1.** Most prevalent integron integrase (IntI) classes

| No. of IntIs in cluster/ class | Prevalence among samples (%) | BLASTP taxa | BLASTP amino acid (%) |
|---|---|---|---|
| 94 | 100 | Class 1 integron - Multispecies | 99.3 |
| 71 | 91.7 | Xanthomonadales (Rhodanobacteraceae and Xanthomonadaceae) | ~70 |
| 19 | 91.7 | Multiple phyla (Deltaproteobacteria, Nitrospinae, Chloroflexi) | ~60 |
| 26 | 66.7 | Xanthomonadaceae (*Lysobacter*, *Vulcaniibacterium*, *Thermomonas*, *Luteimonas*, *Pseudoxanthomonas*) | ~70 |
| 10 | 58.3 | Betaproteobacteria | ~80 |
| 13 | 41.7 | 'IntI1-like' - Multispecies | ~91 |
| 13 | 41.7 | Rhodanobacteraceae | ~75 |
| 12 | 41.7 | Xanthomonadales (Rhodanobacteraceae and Xanthomonadaceae) | ~74 |
| 8 | 41.7 | Xanthomonadales (Rhodanobacteraceae and Xanthomonadaceae) | ~72 |
| 7 | 41.7 | Planctomycetes | ~67 |

### ONT – Illumina comparisons

For comparisons of the cassette and integrase diversity recovered between ONT and Illumina technologies, we first considered differences in sequencing depth. To do this, we randomly selected triplicate 50-megabase (Mb) subsamples from the raw reads of each sample using rasusa v0.3.0 [63] [parameters: --coverage 50 --genome-size 1 Mb]. All sequence processing and filtering steps were repeated for each triplicate subsample as described above.

All formal comparisons were made using two-sample T-tests (or Wilcoxon rank sum tests if the data were not normally distributed) using the rstatix v0.7.0 R package [64]. To determine if the data were normally distributed, Shapiro-Wilk tests were performed using rstatix v0.7.0 [64], and their distributions were visually compared to their theoretical normal distributions using Q-Q plots generated with the R package ggpubr v0.4.0 [65].

To assess the overlap in recovered ORFs between ONT and Illumina, we mapped the cassette ORFs from one technology to the reads of the other using minimap2 v2.20-r1061 [39]. We considered the ORF to be present if it had a mean coverage depth of at least 1 x that spanned at least 98% of the ORF. For ONT and Illumina read mapping, we used the minimap2 presets [-ax map-ont -t 8] and [-ax sr -t 8], respectively. Coverage statistics were extracted from the mapping alignments using the 'sort' and 'coverage' programmes within the SAMtools v1.12 software package [66, 67].

## RESULTS AND DISCUSSION

Here, we present a stringent pipeline for PCR amplifying, sequencing, and analysing integron integrases and gene cassettes from diverse microbial communities (Fig. 1). For this, we used two different PCR primer sets, HS287/HS286 and intI-R / HS286 (Fig. 1a). The sample types consisted of a wide variety of soils (from an urban parkland, an Australian desert, and an Antarctic island), as well as river and estuarine sediments, and freshwater biofilms.

To assess the suitability of long- and short-read sequencing technologies, we sequenced amplicons from both PCR assays using Nanopore (ONT) and Illumina platforms, respectively. The average ONT yield was 181 Mb (100–358 Mb per sample) for the HS287/HS286 primer set and 216 Mb (62–502 Mb per sample) for the intI-R / HS286 primer set. The average Illumina yield was 418 Mb (228–720 Mb per sample) and 663 Mb (275–1247 Mb per sample), respectively for these primer sets.

To ensure amplicons were part of genuine integrons, we filtered the HS287/HS286 data for *attC* sites, and the intI-R / HS286 data for IntI protein sequences based on the IntI-specific additional domain (I2 α-helix) [51, 52] (Fig. 1b). For the HS287/HS286 data, an average of 23.8 and 19.0% of amplicon sequences were retained after filtering for ONT and Illumina, respectively (Fig. S1a, available in the online version of this article). While, for the intI-R / HS286 data, an average of 1.2 and 1.5% of sequences remained after filtering for ONT and Illumina, respectively (Fig. S1b). The difference in proportions of sequences retained after filtering between ONT and Illumina were not statistically significant for either primer set. The low proportion of surviving sequences for the intI-R / HS286 data is likely a result of the intI-R primer binding to other tyrosine recombinases. While many sequences were filtered out, the data retained from this primer set, as described below, include a large, diverse set of both known and entirely novel integron integrases and gene cassettes.

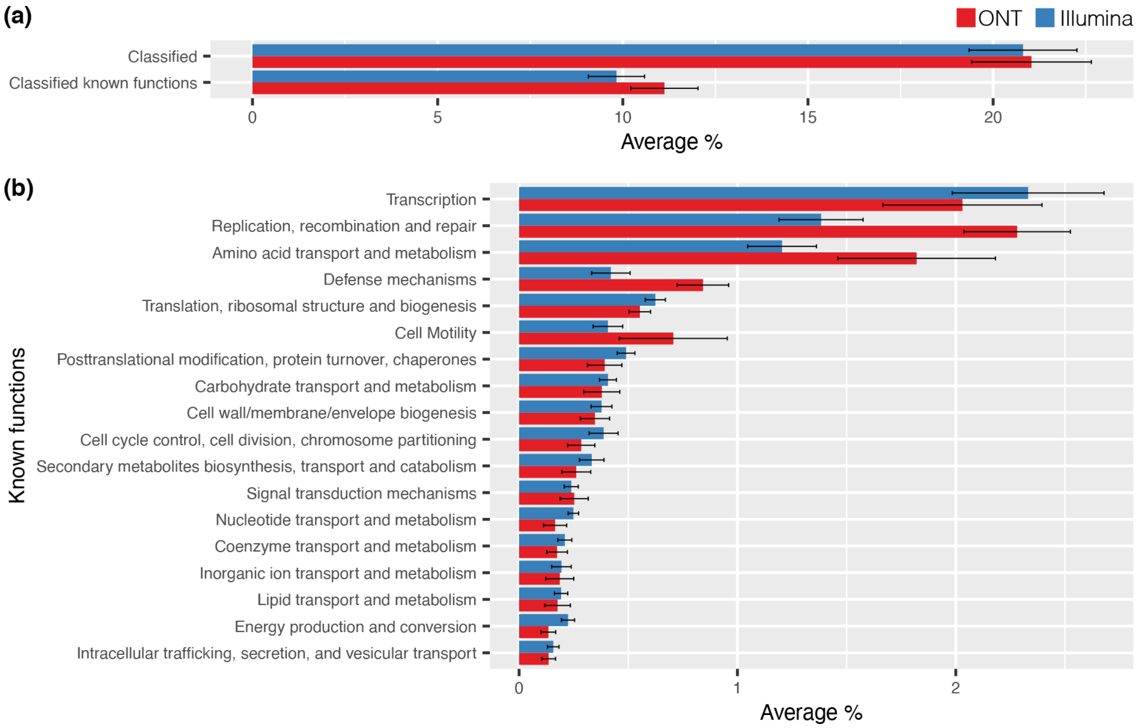

**Fig. 4.** COG functional analysis of cassette-encoded proteins recovered with the HS287/HS286 primer set. (a) Average (±1 S.E) percentage of proteins per sample (*n*=12) that can be classified into functional categories. On average ~20% of cassette-encoded proteins can be classified by a COG category, half of which fall into categories of known function. (b) The average (±1 S.E) proportion of proteins within a sample assigned to each of the known functional categories. The complete list of protein sequences assigned a COG functional category is presented in Table S1.

The lengths of the recovered sequences for both primer sets were significantly larger for ONT sequencing compared to Illumina (Fig. S2). For the HS287/HS286 set, processed sequence lengths ranged from 500 bp to more than 25304 bp for ONT, and 500 to 19244 bp for Illumina. For the intI-R / HS286 data, processed sequence lengths ranged from 803 to 16179 bp for ONT and 800 to 7432 bp for Illumina.

## Recovered diversity of gene cassette ORFs

We assessed the efficiency of both primer sets in recovering gene cassette open reading frames (ORFs). Among all 12 samples, the HS287/HS286 primers amplified 33854 and 62118 non-redundant cassette-encoded proteins when sequenced with ONT and Illumina, respectively (Fig. 2a). After adjusting for sequencing depth, there was no significant difference in cassette recovery between the two sequencing technologies (Fig. 2b). On average, we observed that ~50% of cassette ORFs sequenced with one technology were also recovered by the other (Fig. S3a).

The HS287/HS286 primer set is preferred in order to recover the greatest diversity of gene cassettes. Indeed, the recovery rate of gene cassettes using the methods described here surpasses any previously described approach. Notably, Pereira *et al.* [47] conducted an impressive survey of gene cassettes from 10 terabases of metagenomic data obtained from 14 public databases. Across all datasets, they identified an average of 0.03 unique cassette ORFs per 500 kilobases of assembled data. Here, we recover 218 and 265 ORFs per 500 kilobases of assembled data when sequenced with ONT and Illumina, respectively. Although screening metagenomes may provide a more unbiased approach in analysing gene cassettes, it clearly requires much deeper sequencing to recover sufficient cassette data for in-depth ecological or evolutionary analyses. Studies of integrons and their associated genetic cargo will therefore continue to benefit from the use of amplicon sequencing approaches, such as those described in the present study.

For the intI-R / HS286 primer set, we recovered a total of 9641 and 3742 non-redundant cassette ORFs when sequenced with ONT and Illumina, respectively (Fig. 2c). ONT sequencing recovered a significantly greater number of unique cassette ORFs per sample ($P<0.0001$) than Illumina (Fig. 2d). This was despite the complete alignment of all Illumina reads to the ONT-recovered cassette ORFs (Fig. S3b). The recovery of cassette ORFs from the intI-R/HS286 Illumina reads was likely sub-optimal due to difficulties in the short-read assembly. In particular, different cassette arrays associated with the same or similar integron integrase are likely to be problematic for a short-read assembly approach. While the intI-R / HS286 primer pair does not recover as much diversity

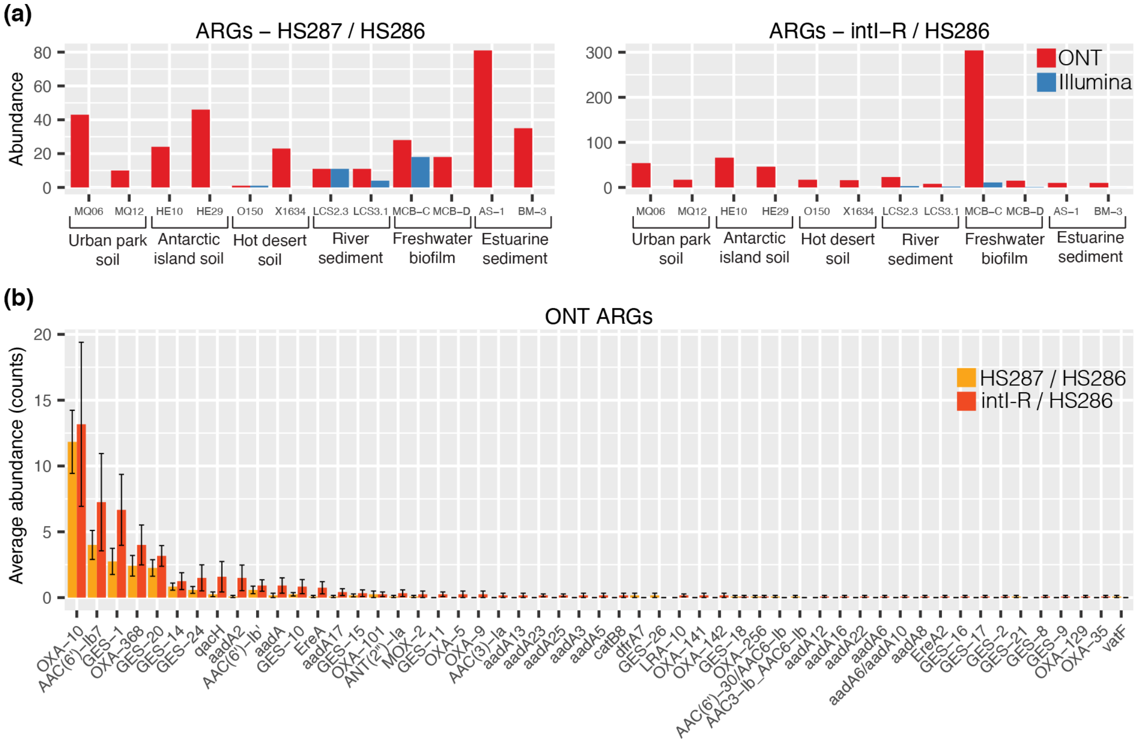

**Fig. 5.** Abundance and diversity of antibiotic resistance gene (ARG) cassettes. (a) Abundance of ARGs recovered from either primer set. (b) The average (±1 S.E) abundance of each ARG type recovered from Nanopore (ONT) sequencing per sample (*n*=12).

as the HS287/HS286 set (*P*<0.0001), it does provide additional key information on IntI diversity (discussed further below) and indicates which gene cassettes are associated with which *intI* genes.

The intI-R / HS286 primer set can also reveal which gene cassettes are located towards the start of a cassette array (Fig. 1a). This is of biological and ecological significance, since the first cassettes in arrays are the most recently inserted cassettes and are likely to be strongly expressed [68]. During environmental perturbations, integron integrase activity leads to the acquisition of novel cassettes, and rearrangement of those already present, inserting them at the start of the array where strong expression is guaranteed [17, 69, 70]. Selection fixes lineages with first-position cassettes that confer significant advantages. Thus, gene cassettes recovered from the intI-R / HS286 primer set might provide important ecological insights at the time of sampling. Such information cannot be obtained from the HS287/HS286 primer set, where cassettes can be amplified from potentially anywhere along a cassette array. Further, the HS287/HS286 primer set will never amplify the first-position cassette of an array, which lack an upstream *attC* site (Fig. 1a).

## Recovered diversity of integron integrases

Using the intI-R / HS286 primers, we recovered a total of 1413 and 1867 different integron integrase genes among the 12 samples when sequenced with ONT and Illumina, respectively (Fig. 3a). There was no significant difference in integron-integrase recovery between the two sequencing technologies, with or without adjusting for sequencing depth (Fig. 3a, b). Both sequencing technologies could recover an impressive diversity of IntIs from the 12 samples. In comparison, a comprehensive screening of 2484 bacterial genomes recovered only 215 different IntIs [13]. This shows that despite the low rate of intI-R / HS286 sequences that are retained after filtering, a significant number of novel integron sequences are recovered.

To determine how many classes of integrons these IntIs represented, we sought to define the amino acid clustering threshold for an integron class. To do this, we used the most abundant and widely distributed IntI, the class 1 integron integrase (IntI1) [71]. Here, we iteratively set decreasing amino acid clustering thresholds for our library of IntIs using CD-HIT v4.6 [72, 73] [parameters: -n 5 -d 0 -g 1 -t 0]. We continued in decreasing increments of 1% until all IntI1s in our dataset were grouped into a single cluster while ensuring all non-IntI1s were excluded (Fig. S4a). This resulted in a 94% amino acid identity being selected as the most appropriate clustering threshold for IntI1s. Although this might not reflect the ideal threshold for all classes, it nevertheless provides a semi-quantitative approach to defining an integron class based on amino acid homology.

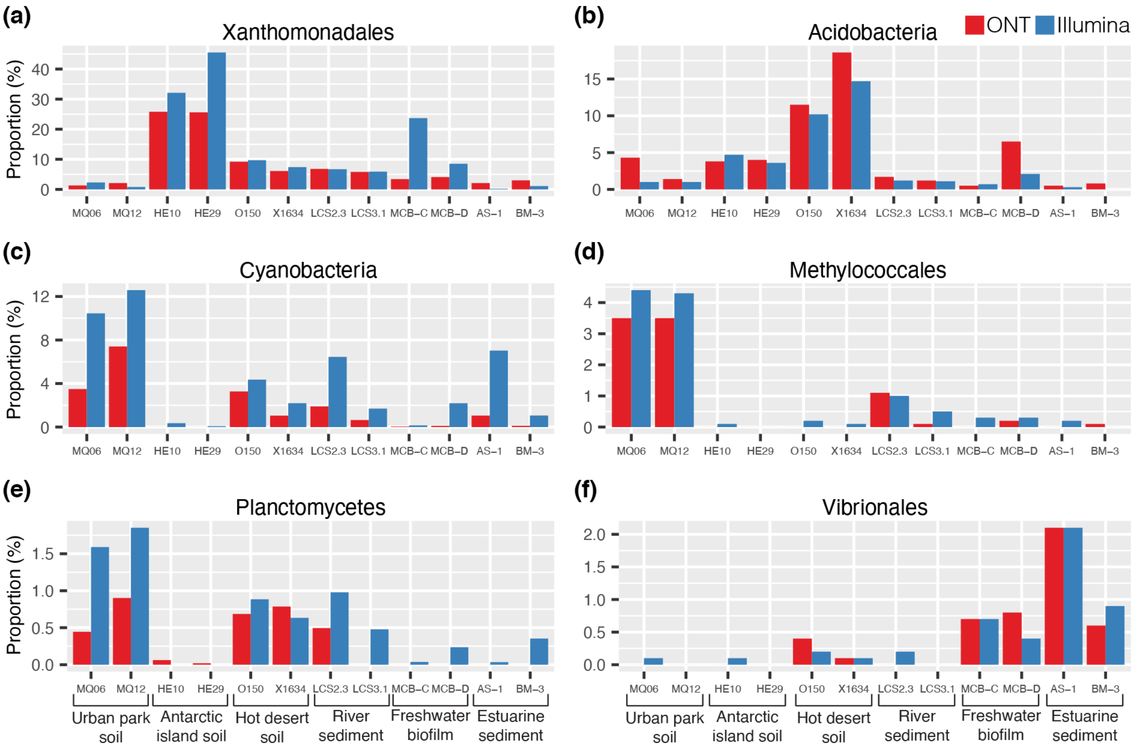

**Fig. 6.** Proportions of gene cassette recombination sites (*attC*s) assigned to bacterial taxa. Taxonomic predictions are based on a selection of six (a–f) of the eleven available taxonomic models of chromosomal *attC*s. Each figure panel shows the proportion of *attC*s across each sample that exhibit sequence and structure conserved among that taxon. For a comparison of all eleven taxa, see Fig. S5.

Using a 94% clustering threshold, we recovered a total of 984 and 1646 integron classes among our dataset when sequenced with ONT and Illumina, respectively (Fig. S4b). There was no significant difference in integron class recovery between the two sequencing technologies, with or without adjusting for sequencing depth (Fig. 3c, d). In addition, we examined the most prevalent integron classes, defined here as IntIs that were present in at least one-third of all samples (Table 1). This identified ten prevalent IntI classes, found to be 60–70% similar to endogenous IntIs from diverse bacterial phyla (Table 1). Not surprisingly, class 1 integrons were the only class to be found in every sample, including those from Antarctica and outback Australia.

## Functional diversity of gene cassettes

Here we show that gene cassette ORFs largely encode proteins of unknown functions (Fig. 4a). This is in agreement with previous functional analyses of gene cassettes [5, 20, 24, 25]. On average, only ~20% of cassette-encoded proteins amplified with HS287/HS286 could be assigned a COG functional category, approximately half of which could be assigned a non-'function unknown' category (Fig. 4a). The dominant COG categories were ‘Transcription’, ‘Replication, recombination and repair’, and ‘Amino acid transport and metabolism’ (Fig. 4b). We show that our methods are capable of recovering gene cassettes that confer a wide range of traits spanning many functional classes.

## Cassette-encoded antibiotic resistance

For a more specific functional characterisation, we focused on antimicrobial resistance, since these phenotypes are often conferred by integron gene cassettes in clinical settings [3, 74, 75]. Interestingly, we found that for either primer set, ONT sequencing could recover many more ARGs than Illumina, the latter recovering no ARGs for most samples (Fig. 5a). In contrast, ONT sequencing recovered as many as 300 ARG cassettes within a single sample. We suspect that this discrepancy is an artefact caused by multiple arrangements of the same ARGs in class 1 cassette arrays that makes their assembly difficult from short-read data. In total, we recovered 106 different ARGs from both primer sets when sequenced with ONT (Fig. 5b). Almost all ARG cassettes encoded proteins known to confer resistance to β-lactam and aminoglycoside antibiotics, these being the most commonly observed integron-mediated resistance types [3].

Upon examining all cassette ORFs associated with class 1 integron integrases recovered using intI-R / HS286 primers (Table 1), we found that 162 of 462 (34.6%) were known ARGs. In comparison, only 586 of the 10385 (5.6%) total cassette ORFs amplified with this primer set were known ARGs. These findings show that class 1 integrons are collecting and concentrating ARG cassettes out of the broader diversity of cassette functions. This enrichment strongly supports the idea that class 1 integrons are key vectors for acquisition and dissemination of antibiotic resistance [3, 76, 77].

### Taxonomic classification of *attC* sites

We could identify the likely taxonomic sources of 5998 *attC*s (18.8%) and 10257 *attC*s (20%) sequenced with ONT and Illumina, respectively. For taxonomic classification, we used models that capture the sequence and structural homology of chromosomal *attC*s from 11 different taxa. These included six Gammaproteobacterial orders (Alteromonadales, Methylococcales, Oceanospirillales, Pseudomonadales, Vibrionales, Xanthomonadales) and an additional five phyla (Acidobacteria, Cyanobacteria, Deltaproteobacteria, Planctomycetes, Spirochaetes). It should be noted that, although the specificity (ability to reject false positives) of each model is very high (98–100%), they exhibit a wide range of sensitivities (proportion of true positives detected) [46]. Therefore, relative abundances of each taxon cannot be compared within the same sample, however, the same taxon can be compared between different samples. This wide range of sensitivities also indicates that the relative abundances of each taxon are likely to be lower-bound estimates.

Here, we show that the relative abundance of each taxon varied across the different sampled environments (Fig. 6). For example, both the Cyanobacteria- and Methylococcales-type *attC*s were most abundant in urban parkland soil (Fig. 6c, d), while Vibrionales-type *attC*s were more abundant in estuarine sediments and freshwater biofilm samples (Figs 6f and S5 for a comparison of all 11 taxa). Such data can provide useful information on the taxonomic contribution to gene cassette pools among different environments.

## CONCLUSIONS

Here, we present experimental and bioinformatic methods for the PCR amplification, DNA sequencing, and analysis of integrons from microbial communities. We describe approaches using two different PCR assays and compare the outputs from ONT and Illumina sequencing. We find that, relative to sequencing depth, ONT generally outperforms or performs the same as Illumina regarding the recovery of gene cassettes and integron integrases. Most notably, ONT outperforms Illumina in the recovery of complete ARG gene cassette sequences. We also find that the primer set HS287/HS286 is efficient at amplifying a wide range of gene cassettes, encompassing extensive *attC* and functional diversity. However, the intI-R / HS286 primer set can provide additional useful information in linking gene cassettes with an integron class. For example, we show that class 1 integrons are collecting and concentrating ARGs relative to the broader cassette pool.

Our described methods can recover key information on the diverse pool of gene cassettes that are helping to drive adaptation and niche specialisation in bacteria [4, 16, 70]. Such an approach allows us to investigate the potential traits that are available to integron-carrying bacteria, and to understand the role that gene cassettes play in mediating evolutionary responses under environmental or clinical selection pressures. In addition, the large proportion of cassettes with unknown functions provides an important resource for the discovery of novel enzymatic activities [17].

Funding information
This research was supported by the Australian Research Council Discovery Grant DP200101874.

Acknowledgements
TMG would like to thank Mary and Saoirse Ghaly for their loving support.

Author contributions
T.M.G., contributed to the conceptualisation of the study and experimental procedures, performed all data analyses, wrote the original draft of the paper, and contributed to the final editing of the manuscript. AnPe., AlPr., Q.Q. and V.R., contributed to the experimental procedures and the final editing of the manuscript. S.G.T. and M.R.G., were responsible for funding acquisition, contributed to the conception of the study and the final editing of the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Gillings MR. Integrons: past, present, and future. *Microbiol Mol Biol Rev* 2014;78:257–277.
2. Mazel D. Integrons: agents of bacterial evolution. *Nat Rev Microbiol* 2006;4:608–620.
3. Partridge SR, Tsafnat G, Coiera E, Iredell JR. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 2009;33:757–784.
4. Gillings MR, Holley MP, Stokes HW, Holmes AJ. Integrons in Xanthomonas: a source of species genome diversity. *Proc Natl Acad Sci U S A* 2005;102:4419–4424.
5. Boucher Y, Labbate M, Koenig JE, Stokes HW. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol* 2007;15:301–309.
6. Stokes HW, Hall RM. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Mol Microbiol* 1989;3:1669–1683.
7. Vit C, Richard E, Fournes F, Whiteway C, Eyer X, *et al*. Cassette recruitment in the chromosomal Integron of *Vibrio cholerae*. *Nucleic Acids Res* 2021;49:5654–5670.
8. Bouvier M, Demarre G, Mazel D. Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J* 2005;24:4356–4367.

9. Loot C, Ducos-Galand M, Escudero JA, Bouvier M, Mazel D. Replicative resolution of integron cassette insertion. *Nucleic Acids Res* 2012;40:8361–8370.

10. Mukhortava A, Pöge M, Grieb MS, Nivina A, Loot C, *et al*. Structural heterogeneity of attC integron recombination sites revealed by optical tweezers. *Nucleic Acids Res* 2019;47:1861–1870.

11. Nivina A, Escudero JA, Vit C, Mazel D, Loot C. Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of attC recombination sites. *Nucleic Acids Res* 2016;44:7792–7803.

12. Chen C-Y, Wu K-M, Chang Y-C, Chang C-H, Tsai H-C, *et al*. Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res* 2003;13:2577–2587.

13. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* 2016;44:4539–4550.

14. Cambray G, Sanchez-Alberola N, Campoy S, Guerin É, Da Re S, *et al*. Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mob DNA* 2011;2:6.

15. Guerin E, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, *et al*. The SOS response controls integron recombination. *Science* 2009;324:1034.

16. Escudero JA, Loot C, Nivina A, Mazel D. The integron: adaptation on demand. *Microbiol Spectr* 2015;3:MDNA3-0019.

17. Ghaly TM, Geoghegan JL, Tetu SG, Gillings MR. The peril and promise of integrons: beyond antibiotic resistance. *Trends Microbiol* 2020;28:455–464.

18. Ghaly TM, Chow L, Asher AJ, Waldron LS, Gillings MR. Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS One* 2017;12:e0179169.

19. Gillings MR. Class 1 integrons as invasive species. *Curr Opin Microbiol* 2017;38:10–15.

20. Ghaly TM, Geoghegan JL, Alroy J, Gillings MR. High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environ Microbiol* 2019;21:1567–1574.

21. Elsaied H, Stokes HW, Nakamura T, Kitamura K, Fuse H, *et al*. Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environ Microbiol* 2007;9:2298–2312.

22. Elsaied H, Stokes HW, Kitamura K, Kurusu Y, Kamagata Y, *et al*. Marine integrons containing novel integrase genes, attachment sites, attI, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *ISME J* 2011;5:1162–1177.

23. Elsaied H, Stokes HW, Kitamura K, Kurusu Y, Kamagata Y, *et al*. Marine integrons containing novel integrase genes, attachment sites, *attI*, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *ISME J* 2011;5:1162–1177.

24. Koenig JE, Boucher Y, Charlebois RL, Nesbø C, Zhaxybayeva O, *et al*. Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol* 2008;10:1024–1038.

25. Koenig JE, Sharp C, Dlutek M, Curtis B, Joss M, *et al*. Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney tar ponds. *PLoS One* 2009;4:e5276.

26. Dias MF, de Castro GM, de Paiva MC, de Paula Reis M, Facchin S, *et al*. Exploring antibiotic resistance in environmental integron-cassettes through intI-attC amplicons deep sequencing. *Braz J Microbiol* 2021;52:363–372.

27. Ghaly TM, Gillings MR, Penesyan A, Qi Q, Rajabal V, *et al*. The natural history of integrons. *Microorganisms* 2021;9:11.

28. Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KM, *et al*. Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol* 2001;67:5240–5246.

29. Green JL, Holmes AJ, Westoby M, Oliver I, Briscoe D, *et al*. Spatial scaling of microbial eukaryote diversity. *Nature* 2004;432:747–750.

30. Oliver I, Holmes A, Dangerfield JM, Gillings M, Pik AJ, *et al*. Land systems as surrogates for biodiversity in conservation planning. *Ecological Applications* 2004;14:485–503.

31. Gillings MR, Krishnan S, Worden PJ, Hardwick SA. Recovery of diverse genes for class 1 integron-integrases from environmental DNA samples. *FEMS Microbiol Lett* 2008;287:56–62.

32. Yeates C, Gillings MR, Davison AD, Altavilla N, Veal DA. Methods for microbial DNA extraction from soil for PCR amplification. *Biol Proced Online* 1998;1:40–47.

33. Chow L, Waldron L, Gillings MR. Potential impacts of aquatic pollutants: sub-clinical antibiotic concentrations induce genome changes and promote antibiotic resistance. *Front Microbiol* 2015;6:803.

34. Guppy v4.3.4: Local accelerated basecalling for Nanopore data; https://community.nanoporetech.com/downloads

35. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.

36. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al*. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–736.

37. Bushnell B. BBTools software package. 579; 2014. http://sourceforge.net/projects/bbmap

38. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* 2017;27:737–746.

39. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.

40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.

42. Deng M, Jiang R, Sun F, Zhang X. Research in computational molecular biology. In: A K (eds). *Assembling Genomes and Mini-Metagenomes from Highly Chimeric Reads*. Berlin, Heidelberg: Springer; 2013.

43. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics* 2020;70:e102.

44. Pereira MB, Wallroth M, Kristiansson E, Axelson-Fisk M. HattCI: fast and accurate *attC* site identification using hidden markov models. *J Comput Biol* 2016;23:891–902.

45. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933–2935.

46. Ghaly TM, Tetu SG, Gillings MR. Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of attC sites. *Commun Biol* 2021;4:946.

47. Buongermino Pereira M, Österlund T, Eriksson KM, Backhaus T, Axelson-Fisk M, *et al*. A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* 2020;21:1–14.

48. Marijon P, Chikhi R, Varré J-S. yacrd and fpa: upstream tools for long-read genome assembly. *Bioinformatics* 2020;36:3894–3896.

49. Sahlin K, Medvedev P. *De novo* clustering of long-read transcriptome data using a greedy, quality value-based algorithm. *J Comput Biol* 2020;27:472–484.

50. Sahlin K, Medvedev P. Author Correction: Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat Commun* 2021;12:992.

51. Messier N, Roy PH. Integron integrases possess a unique additional domain necessary for activity. *J Bacteriol* 2001;183:6699–6706.

52. Escudero JA, Nivina A, Kemble HE, Loot C, Tenaillon O, *et al*. Primary and promiscuous functions coexist during evolutionary innovation through whole protein domain acquisitions. *elife* 2020;9:e58061.

53. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.

54. Eddy SR. HMMER 3.2: Biosequence analysis using profile hidden Markov models; 2018. http://hmmer.org/
55. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–36.
56. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* 2017;34:2115–2122.
57. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
58. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
59. Seemann T. *ABRicate: Mass Screening of Contigs for Antimicrobial and Virulence Genes*. Melbourne, Australia: Department of Microbiology and Immunology, The University of Melbourne; 2018.
60. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, *et al*. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;45:D566–D573.
61. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, Mazel D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res* 2003;13:428–442.
62. Vaisvila R, Morgan RD, Posfai J, Raleigh EA. Discovery and distribution of super-integrons among pseudomonads. *Mol Microbiol* 2001;42:587–601.
63. Hall MB. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J Open Source Softw* 2019.
64. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0; 2021. https://CRAN.R-project.org/package=rstatix
65. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots. R package version 0.4.0; 2020. https://CRAN.R-project.org/package=ggpubr
66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
67. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10:giab008.
68. Collis CM, Hall RM. Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* 1995;39:155–162.
69. Souque C, Escudero JA, MacLean RC. Integron activity accelerates the evolution of antibiotic resistance. *Elife* 2021;10:e62474.
70. Escudero JA, Loot C, Mazel D. Integrons as adaptive devices. In: Rampelotto PH (eds). *Molecular Mechanisms of Microbial Evolution*. Springer International Publishing; 2018. pp. 199–239.
71. Zhu Y-G, Gillings M, Simonet P, Stekel D, Banwart S, *et al*. Microbial mass movements. *Science* 2017;357:1099–1100.
72. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;17:282–283.
73. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 2002;18:77–82.
74. Zhu Y-G, Zhao Y, Li B, Huang C-L, Zhang S-Y, *et al*. Continental-scale pollution of estuaries with antibiotic resistance genes. *Nat Microbiol* 2017;2:16270.
75. Ghaly TM, Paulsen IT, Sajjad A, Tetu SG, Gillings MR. A novel family of *Acinetobacter* mega-plasmids are disseminating multidrug resistance across the globe while acquiring location-specific accessory genes. *Front Microbiol* 2020;11:3058.
76. Rowe-Magnus DA, Mazel D. The role of integrons in antibiotic resistance gene capture. *Int J Med Microbiol* 2002;292:115–125.
77. Stalder T, Barraud O, Casellas M, Dagot C, Ploy M-C. Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol* 2012;3:119.