





Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of *attC* sites

Timothy M. Ghaly ¹, Sasha G. Tetu ^{2,3} & Michael R. Gillings ^{1,3}

Integrans are bacterial genetic elements that can capture mobile gene cassettes. They are mostly known for their role in the spread of antibiotic resistance cassettes, contributing significantly to the global resistance crisis. These resistance cassettes likely originated from sedentary chromosomal integrans, having subsequently been acquired and disseminated by mobilised integrans. However, their taxonomic and environmental origins are unknown. Here, we use cassette recombination sites (*attCs*) to predict the origins of those resistance cassettes now spread by mobile integrans. We modelled the structure and sequence homology of 1,978 chromosomal *attCs* from 11 different taxa. Using these models, we show that at least 27% of resistance cassettes have *attCs* that are structurally conserved among one of three taxa (Xanthomonadales, Spirochaetes and Vibrionales). Indeed, we found some resistance cassettes still residing in sedentary chromosomal integrans of the predicted taxa. Further, we show that *attCs* cluster according to host environment rather than host phylogeny, allowing us to assign their likely environmental sources. For example, the majority of β -lactamases and aminoglycoside acetyltransferases, the two most prevalent resistance cassettes, appear to have originated from marine environments. Together, our data represent the first evidence of the taxonomic and environmental origins of resistance cassettes spread by mobile integrans.

¹Department of Biological Sciences, Macquarie University, Sydney, Australia. ²Department of Molecular Sciences, Macquarie University, Sydney, Australia. ³ARC Centre of Excellence in Synthetic Biology, Macquarie University, Sydney, Australia. ✉email: timothy.ghaly@mq.edu.au

Integrations are bacterial genetic elements that can insert and excise mobile gene cassettes by site-specific recombination. They are mostly known for the accumulation and spread of antibiotic resistance gene cassettes among Gram-negative bacteria¹, although it is clear that they can mobilise and rearrange a diverse range of cassettes that encode functions well beyond clinical relevance². The recruitment of gene cassettes by an integron is mediated by its core functional gene, the integron integrase (IntI). Once inserted, gene cassettes form part of a cassette array that can vary considerably in size, ranging from one or two cassettes to more than three hundred^{3–5}. Cassette arrays are highly variable regions, in which the content and arrangement of genes can be changed by IntI activity, often induced by environmental stress^{6,7}. Consequently, integrons provide genomic plasticity and adaptation on demand⁸.

Integrations have been classified into two main groups. The first are mobile integrons, which comprise five known IntI classes that have become embedded into mobile elements. These integrons, particularly the class 1 integron, are of clinical significance, often harbouring multiple antibiotic-resistance cassettes. The class 1 integron is believed to have become mobilised from a Betaproteobacterial chromosome in the early 20th century^{1,9}. Since that time, derivatives of this ancestral element have spread into more than 70 clinically important bacteria and have acquired more than 130 different resistance genes that confer resistance to most classes of antibiotics^{10,11}. Mobile class 1 integrons have been found on every continent, including Antarctica, and are ubiquitous in the microbiomes of humans and agricultural animals^{11,12}. This successful colonisation means that up to 10^{23} copies of class 1 integrons are shed into the environment every day via human and agricultural waste¹².

Class 1 integrons, together with the other four mobile classes, vector almost all antibiotic resistance gene cassettes that can be detected in bacterial genomes⁵. Consequently, their success has been largely driven by strong antimicrobial selection. They differ considerably from sedentary chromosomal integrons (SCIs), which represent the ancestral state of integrons¹³. These largely carry gene cassettes of unknown functions², and are present in ~17% of sequenced genomes¹⁴.

Gene cassettes are mobilisable elements that carry a cassette-associated recombination site (*attC*). The *attC* site allows insertion into a cassette array at the integron-associated recombination site (*attI*) (Fig. 1A). Insertion involves the recombination between *attI* and only the bottom strand of the *attC* site^{15,16}, forming an atypical Holliday junction (Fig. 1B). As only one strand of the *attC* is involved in recombination, the Holliday junction cannot be resolved by the typical second-strand exchange. Instead, it is resolved by replication of the entire recombinant molecule¹⁷ (Fig. 1B). Folding of the *attC* single strand is permitted by the pairing of two sets of inverted repeats (*R'* to *R''* and *L'* to *L''*) (Fig. 1C). For most *attCs*, however, the number of nucleotides involved in pairing extends beyond the R and L boxes, which influences the final *attC* secondary structure. As a consequence, the overall structure and length of *attCs* can vary considerably between different gene cassettes (Fig. 1D).

The efficiency of recombination is largely dependent on the folded hairpin structure of the bottom *attC* strand^{18–20}. Further, different integron integrases can recognise different ranges of *attC* structures. In particular, the class 1 integron integrase (IntI1) has a broad *attC* specificity range^{21,22}. This is likely to be one of the reasons for its remarkable success after it became mobilised, because it can gain access to, and recognise, diverse gene cassettes from broad phylogenetic sources. Indeed, cassettes associated with mobile class 1 integrons appear to have been acquired from diverse chromosomal contexts. This is evident from the inconsistent codon usage in their cassette open reading frames (ORFs)

and the considerable sequence and structural diversity of their *attC* sites¹⁴. In contrast, SCIs generally contain cassettes with highly similar *attCs*, and these share homology with other *attCs* in the same bacterial taxon^{23–25}. This conservation provides an opportunity to predict the taxonomic origins of clinically important gene cassettes disseminated by mobile integrons. Hereafter, we use the term ‘origin’ to represent the SCI source from which the cassettes were acquired by mobile integrons.

Knowledge of the taxonomic origins of resistance cassettes, and what ecological and physiological traits are shared by these taxa, might allow us to predict environmental hotspots or conditions that contribute to the emergence of novel resistance cassettes²⁶. This could suggest efficient mitigation strategies to prevent the spread of novel resistance genes into clinical settings.

Here, we modelled the conserved structure and sequence homology of *attCs* from distinct bacterial taxa. We used these taxon-specific models to predict the origins of resistance gene cassettes found on mobile integrons and provide evidence for some of these cassettes still residing in the SCIs of those taxa. Further, we show that *attCs* are more similar among bacteria that inhabit a similar environment, allowing us to predict the environmental sources of each resistance cassette. In particular, we show that the majority of β -lactamases and aminoglycoside acetyltransferases, the two most prevalent resistant cassettes, appear to have originated in marine bacteria. We also find that both the structure of *attCs* and the phylogeny of IntIs cluster according to the host environment rather than host phylogeny. We propose that this shared clustering pattern is the result of convergent and co-evolutionary processes. IntI-*attC* co-evolution would allow recombination specificity to be maintained within a taxon, while convergent evolution facilitates the successful exchange of cassettes between divergent taxa inhabiting the same environment.

Results and discussion

Efficacy of *attC* taxonomy predictions. We used the sequence and structural homology of *attC* recombination sites, conserved within individual taxa, to predict the sources of gene cassettes spread by mobile integrons. To do this, we generated covariance models (CMs) built from sets of taxon-specific *attCs* obtained from the chromosomes of diverse bacteria. CMs are similar to profile hidden Markov models (HMMs) in that they both capture position-specific information about how conserved each column of an alignment is, and which nucleotides/residues are likely to occur. However, in a profile HMM, each position of the profile is treated independently, while in a CM, base-paired positions (when folded) are dependent on one another, therefore, modelling the covariation at these positions. This is necessary to assess correct base-pairing in secondary structure formation. Thus, both conserved sequence and secondary structure of taxon-specific *attCs* can be modelled.

We determined the specificity and sensitivity of each model in assigning the correct taxonomy to our complete *attC* dataset ($n = 2,352$). The efficacy of each CM was determined by its sensitivity (capacity to detect *attCs* from the taxon that it was built from), and its specificity (ability to exclude *attCs* derived from other taxa). CMs that did not achieve a specificity greater than 98% were excluded. This resulted in a set of 11 taxon-specific CMs that proved efficient in the taxonomic assignment of *attCs* (Supplementary Table 1). The taxa comprised of six Gammaproteobacterial orders (Alteromonadales, Methylococcales, Oceanospirillales, Pseudomonadales, Vibrionales, Xanthomonadales) and an additional five phyla (Acidobacteria, Cyanobacteria, Deltaproteobacteria, Planctomycetes, Spirochaetes). In total, 1978 chromosomal *attCs* were used to generate the 11 CMs, ranging from 51–505 *attCs* used for each. Bit score cut-offs for the taxonomic assignment were set for each CM individually in order to

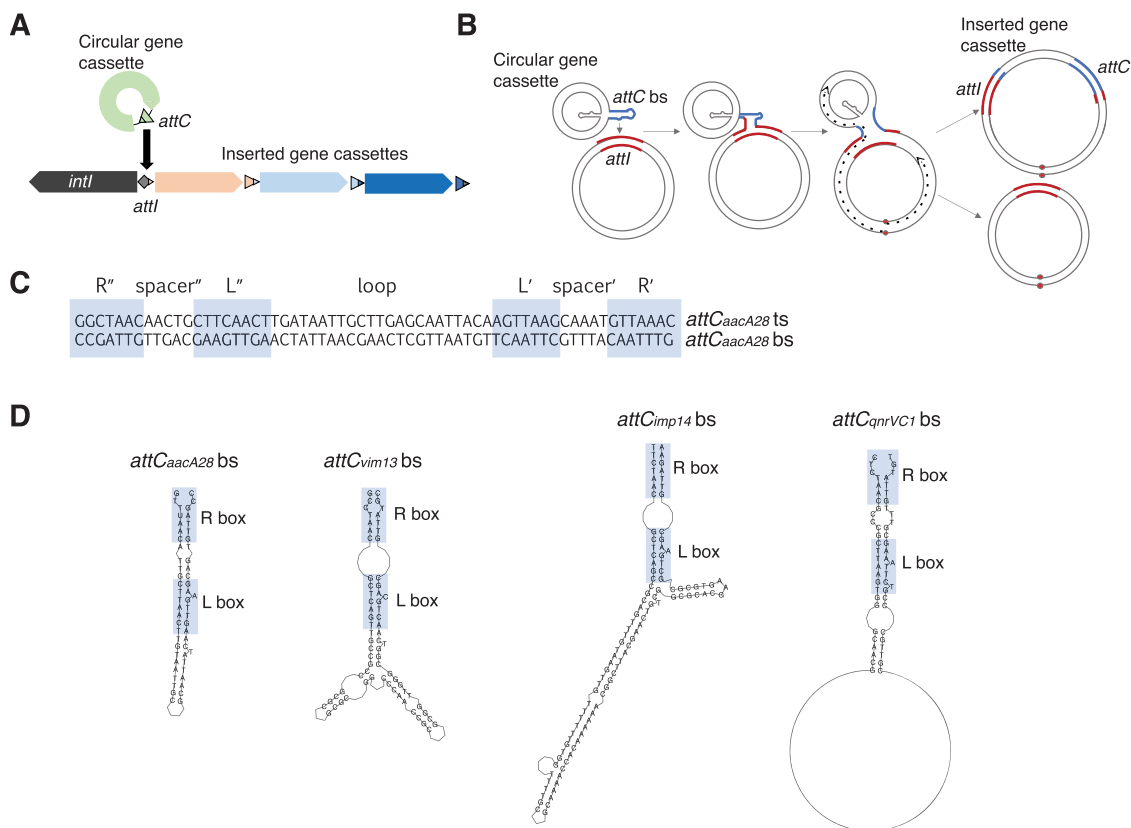


Fig. 1 The role of *attC* folding structure in gene cassette insertion. **A** Integrons carry an integron integrase gene (*intI*) that encodes a tyrosine recombinase (IntI). IntI facilitates the insertion of circular gene cassettes by mediating recombination between the cassette-associated recombination (*attC*) and integron-associated recombination (*attI*) sites. IntI activity can result in arrays of gene cassettes that vary considerably in size (1 to +300). **B** Cassette insertion involves the recombination between *attI* and only the bottom strand of *attC* (*attC* bs). This results in an atypical Holliday junction, which can only be resolved by replication (dotted black arrows; lagging strand not shown). Replication of the recombinogenic strand produces a daughter molecule with the inserted cassette at the *attI* site, while replication of the alternate strand generates the integron without the inserting cassette. **C** The palindromic nature of *attCs* gives rise to their single-stranded folding structure. All *attC* sites have two sets of inverted repeats (R'/R' and L'/L'), which allow the folding of single-stranded *attCs*. Two spacers, spacer'' and spacer', separate R'' from L'' and L' from R', respectively. The middle region of the *attC* is known as the loop and is highly variable in sequence and size. **D** Shown are the predicted bottom strand folding structures of *attCs* from four antibiotic resistance gene cassettes (*aacA28*, *vim13*, *imp14*, and *qnrVC1*). The variable degree of base-pairing beyond the R and L boxes generates considerable structural diversity among different *attCs*, which in turn impacts their recombination efficiency by different IntIs. Folding structures were predicted by RNAfold v 2.4.16 from the ViennaRNA Package 2.0⁴⁰.

maximise sensitivity while ensuring specificity was 98–100% (Supplementary Table 1). The mean specificity for the CMs was 99.6%, ranging from 98.04–100%, and the mean sensitivity was 66.1%, ranging from 30–99.2%. The wide range of CM sensitivities means that the relative number of matches to each taxon cannot be compared, as some have a higher capacity to detect true positives than others. It does, however, indicate that the number of matches to a particular taxon is likely to be a lower-bound estimate.

Predicting taxonomic origins of resistance gene cassettes from mobile integrons. Mobile integrons can transfer between species, allowing the acquisition of new gene cassettes from diverse genomic backgrounds²⁷. Here, we aimed to predict the original taxonomy of resistance gene cassettes commonly found on mobile integrons using the structural and sequence homology of their *attC* recombination sites.

We used a collection of 108 resistance gene cassettes, reported by Partridge et al.¹⁰. Using our models, we found that at least 27% of these had *attCs* structurally conserved among one of three taxa (Table 1). On this basis, nineteen resistance cassettes were assigned to Xanthomonadales, six to Spirochaetes, and four to

Vibrionales. It is important to note, however, that each CM exhibits different sensitivities in their ability to detect true positives (Supplementary Table 1), thus the relative contribution of each taxa to the pool of resistance cassettes cannot be compared and all are likely to be lower-bound estimates. Nevertheless, these three taxa have contributed to more than a quarter of resistance cassettes in our dataset, signifying that they are key taxonomic sources of the resistance cassettes now circulating among diverse Gram-negative pathogens.

The types of resistance mechanisms varied between taxa (Table 1). For example, cassettes with Xanthomonadales-type *attCs* encoded a wide range of different resistance proteins, consisting of aminoglycoside (3'') adenylyltransferases, aminoglycoside (2'') adenylyltransferases, aminoglycoside (6') acetyltransferases, aminoglycoside (3') acetyltransferases, chloramphenicol acetyltransferases, small drug resistance proteins, a QAC efflux pump, a streptothricin acetyltransferase, and a fosfomycin resistance protein. Whilst, the six cassettes with Spirochaetes-type *attCs* all encoded aminoglycoside (6') acetyltransferases. Interestingly, cassettes with a predicted Spirochaetes origin represent 30% of all aminoglycoside (6') acetyltransferases in our dataset, highlighting this phylum as a significant source of this resistance mechanism.

Table 1 Predicted taxonomic source of resistance gene cassettes found on mobile integrons.

Source taxon	Gene cassette	Gene cassette product	Example recipient host from ref. ¹⁰	Accession from ref. ¹⁰
Xanthomonadales	<i>aadA1a</i>	Aminoglycoside (3") adenylyltransferase	<i>Escherichia coli</i>	X12870.1
Xanthomonadales	<i>aadA7</i>	Aminoglycoside (3") adenylyltransferase	<i>Escherichia coli</i>	AF224733.1
Xanthomonadales	<i>qacI</i>	Quaternary ammonium compound efflux protein	<i>Escherichia coli</i>	AF205943.1
Xanthomonadales	<i>aacA3</i>	Aminoglycoside (6') acetyltransferase	<i>Salmonella enterica</i>	AY123251.1
Xanthomonadales	<i>catB3</i>	Chloramphenicol acetyltransferase	<i>Enterobacter aerogenes</i>	U13880.2
Xanthomonadales	<i>qacF</i>	Quaternary ammonium compound efflux protein	<i>Enterobacter aerogenes</i>	AF034958.3
Xanthomonadales	<i>aadA6</i>	Aminoglycoside (3") adenylyltransferase	<i>Pseudomonas aeruginosa</i>	AF140629.1
Xanthomonadales	<i>aadA24</i>	Aminoglycoside (3") adenylyltransferase	<i>Salmonella enteritidis</i>	AM711129.1
Xanthomonadales	<i>aacA37</i>	Aminoglycoside (6') acetyltransferase	<i>Pseudomonas aeruginosa</i>	DQ302723.1
Xanthomonadales	<i>catB5</i>	Chloramphenicol acetyltransferase	<i>Morganella morganii</i>	X82455.1
Xanthomonadales	<i>aadB</i>	Aminoglycoside (2") adenylyltransferase	<i>Escherichia coli</i>	L06418.4
Xanthomonadales	<i>aadA2</i>	Aminoglycoside (3") adenylyltransferase	<i>Escherichia coli</i>	X68227.1
Xanthomonadales	<i>sat2</i>	Streptothricin acetyltransferase	<i>Escherichia coli</i>	X15995.1
Xanthomonadales	<i>fosE</i>	Fosfomycin resistance protein	<i>Pseudomonas aeruginosa</i>	AY029772.1
Xanthomonadales	<i>aacA7</i>	Aminoglycoside (6') acetyltransferase	<i>Enterobacter aerogenes</i>	U13880.2
Xanthomonadales	<i>aacC6</i>	Aminoglycoside (3) acetyltransferase	<i>Serratia marcescens</i>	AY884051.1
Xanthomonadales	<i>smr2</i>	Small multidrug resistance protein	<i>Escherichia coli</i>	AY260546.3
Xanthomonadales	<i>aacA29</i>	Aminoglycoside (6') acetyltransferase	Uncultured bacterium	AY139599.1
Xanthomonadales	<i>aacC5</i>	Aminoglycoside (3) acetyltransferase	<i>Vibrio fluvialis</i>	AB114632.1
Spirochaetes	<i>aacA39</i>	Aminoglycoside (6') acetyltransferase	<i>Pseudomonas aeruginosa</i>	EU886977.1
Spirochaetes	<i>aacA1:qcuG</i>	Aminoglycoside (6') acetyltransferase	<i>Escherichia coli</i>	AF047479.2
Spirochaetes	<i>aacA30</i>	Aminoglycoside (6') acetyltransferase	<i>Salmonella enterica</i>	AY289608.1
Spirochaetes	<i>aacA17</i>	Aminoglycoside (6') acetyltransferase	<i>Klebsiella pneumoniae</i>	AF047556.1
Spirochaetes	<i>aacA16</i>	Aminoglycoside (6') acetyltransferase	<i>Citrobacter freundii</i>	Z54241.1
Spirochaetes	<i>aacA28</i>	Aminoglycoside (6') acetyltransferase	<i>Pseudomonas aeruginosa</i>	AB104852.1
Vibrionales	<i>dfrA6</i>	Dihydrofolate reductase	<i>Proteus mirabilis</i>	Z86002.1
Vibrionales	<i>blaP3</i>	Class A β -lactamase	<i>Pseudomonas aeruginosa</i>	U14749.1
Vibrionales	<i>qnrVC1</i>	Quinolone resistance protein	<i>Vibrio cholerae</i>	EU436855.2
Vibrionales	<i>blaP7</i>	Class A β -lactamase	<i>Vibrio cholerae</i>	AF409092.1

To further validate our findings, we searched for examples of the complete cassettes in the predicted taxon of origin. To do this, we searched for the cassette open reading frames (ORFs) among all complete bacterial chromosomes available in NCBI ($n = 24,143$ as of 27th January 2021). We found that three of the resistance cassette ORFs were present in bacterial chromosomes, and in all cases, they were part of gene cassettes. All three were only found in the taxa predicted by our CMs and were only present once each among all 24,143 chromosomes. All three were part of gene cassettes within SCIs (100% nucleotide identity

spanning the ORF and *attC* site). Two of the cassettes, which we had predicted to originate from Xanthomonadales, encoded a chloramphenicol acetyltransferase (*catB3*) and a QAC efflux pump (*qacI*). Interestingly, both of these were found in the same chromosomal integron of *Lysobacter oculi* (Accession: CP029556.1), part of the order Xanthomonadales. The third cassette encoded a quinolone resistance protein (*qnrVC1*) that we predicted to have a Vibrionales origin, and was indeed found within a cassette array of *Vibrio alginolyticus* (Accession: CP060386.1). The fact that all of these ORFs were found in the

taxa predicted by our CMs, and were not present in any other taxa, strongly supports the predictions made by our CMs. In addition, the *attC* of the *blaP3* resistance cassette, which our modelling predicted to have originated from Vibrionales has previously been reported to share overall sequence similarity of 90% with *Vibrio cholerae attC* sequences²⁸.

Knowledge of the taxonomic origins of resistance gene cassettes, and what ecological and physiological traits are shared by these taxa, can allow us to predict environmental hotspots or conditions that contribute to the emergence of novel resistance genes²⁶. The origins of up to 30 antibiotic resistance genes have been previously proposed, however, none of these attributions included integron gene cassettes²⁶. Predicting chromosomal origins of resistance genes generally involves examining the genes in the immediate vicinity that have been co-mobilised²⁶. However, this approach cannot work for integron gene cassettes, as each cassette is a modular mobile unit, thus, co-mobilisation and continuous maintenance of the same arrangement of multiple cassettes is extremely unlikely. However, by using the taxonomic signatures preserved in the structure and sequence of their recombination sites, we are able to predict from which taxa cassettes have likely originated.

Environmental clustering of *attC*s: predicting environments of origin of resistance cassettes. Several studies have shown that the phylogeny of integron integrases (IntI) cluster according to the host environment, with marine IntIs forming one major clade and soil/freshwater IntIs forming another^{4,6,29}. The inverted IntIs form a sub-clade within the larger soil/freshwater clade²⁹. We therefore investigated whether *attC*s exhibit similar environmentally explicit clustering. To do this, we used ten representative *attC*s from each taxon-specific CM, which together with the 108 resistance cassettes from mobile integrons¹⁰, were clustered based on sequence and folding structure.

We found that *attC*s cluster into three major clades (Fig. 2). One that is distinctly a marine clade, the second is a soil/freshwater clade, and a third clade that we have labelled as ‘Xanthomonadales-like’ (XL), as only *attC*s from the Xanthomonadales CM fell into this clade. These distinct clusters allow us to infer from which environments resistance cassettes have likely originated. The XL clade included the greatest number of resistance cassette *attC*s and encompassed the greatest range of resistance types (Fig. 2). However, the majority of β -lactamases (52.9%) and aminoglycoside acetyltransferases (54.2%), the two most common types of antibiotic resistance cassettes¹⁰, formed part of the marine clade. We therefore, for the first time, highlight the marine environment as a key source of these two prevalent types of integron-mediated resistance.

To further investigate if any other taxa carry Xanthomonadales-like *attC*s, we built a CM using the 58 *attC*s from the Xanthomonadales-like clade (Fig. 2) and used this to search all 24,143 bacterial chromosomes available in NCBI. We detected 941 Xanthomonadales-like *attC*s within 56 genomes (Supplementary Table 2). These were almost all found in Xanthomonadales, with 935 *attC*s (99.4%) from 54 Xanthomonadales genomes. Among them were the genera *Xanthomonas*, *Lysobacter*, *Luteimonas*, *Pseudoxanthomonas*, *Thermomonas*, and *Stenotrophomonas* (Supplementary Table 3). Most of these genera are commonly found associated with plant leaves and roots^{30–33}, suggesting that the Xanthomonadales-like *attC* clade might represent plant-associated environments. If this is the case, then plant-associated bacteria might be a significant source of a large proportion of resistance gene cassettes. Interestingly, the class 1 integron platform is also thought to have originated from plant leaf surfaces⁹. However,

since not all of these Xanthomonadales genera are endemic to plants and since we lack additional taxa outside of Xanthomonadales, and thus cannot distinguish if this clade represents an environmental or taxonomic grouping, further evidence would be required to support this.

Convergent and co-evolution of *attC* sites and integron integrases. Here, we show that *attC*s cluster according to the host environment, with a clear distinction between marine *attC*s and soil/freshwater *attC*s. The same pattern has been previously shown for the phylogeny of integron integrases^{4,6,29}. Using IntI protein sequences extracted from genomes used in the present study, we also show that their phylogeny clusters according to the host environment (Supplementary Figure 2). We propose that this shared clustering pattern is the result of convergent and co-evolutionary processes.

Cassette insertion and excision are driven by IntI-mediated *attC* x *attI* and *attC* x *attC* recombination, respectively. The efficiency of these recombinations largely depends on the folding structure of *attC*s¹⁸ (Fig. 1D). However, the recombination efficiency over a range of diverse *attC* sites varies between different integron integrases. For example, the sedentary *Vibrio cholerae* IntIA can recognise a narrower range of *attC*s than those recognised by the mobile class 1 integron integrase²¹. Thus, within a bacterial taxon, the degree of divergence in *attC* structures will likely predict the *attC* specificity of the endogenous integrase. As the *intI* gene evolves, the *attC* folding structures will likely co-evolve to maintain functionality, and vice versa. Additionally, cassette genesis will also influence the degree of divergence among *attC*s. Given the variability in *attC* homogeneity within a taxon, it is possible that some species might generate highly similar *attC*s, while other species might generate variable *attC*s. In either scenario, however, the native mechanism of cassette genesis, along with the selection strength that controls *attC* divergence, and the specificity range of the endogenous integrase are all intrinsically linked. Co-evolution of these three processes would ensure the maintenance of integron functionality and recombination efficiency.

Further, horizontal transfer of cassettes is likely to be more prevalent among bacteria inhabiting the same environment. Here, selection will favour integrons that can recognise and incorporate foreign cassettes with novel functions. This is strongly supported by the observed environmental clustering of *attC* structures and IntI phylogeny. Further, the clustering of IntIs and *attC*s are incongruent with the phylogenetic clustering of their host organisms³⁴. Thus, taxa that are not necessarily closely related by descent, have more similar *intI* genes and *attC* structures. This is evidence of convergent evolution, allowing IntI recombinatorial activity on structurally similar *attC* substrates that are more prevalent in the same environment. A selective advantage would be gained by integron platforms that can successfully recognise and incorporate foreign cassettes from the local environment.

Conclusion

In the present study, we show for the first time that the taxonomic origins of integron gene cassettes can be predicted from the structure and sequence of their *attC* recombination sites. Using this approach, we predicted the source taxon of 29 out of 108 resistance gene cassettes spread by mobile integrons. These appear to have originated from three taxa (Xanthomonadales, Spirochaetes, and Vibrionales). We searched for evidence of these cassettes residing in their presumptive ancestral hosts. We found three of these cassettes in chromosomal integrons in the predicted taxon, and these cassettes were not present in any other bacterial

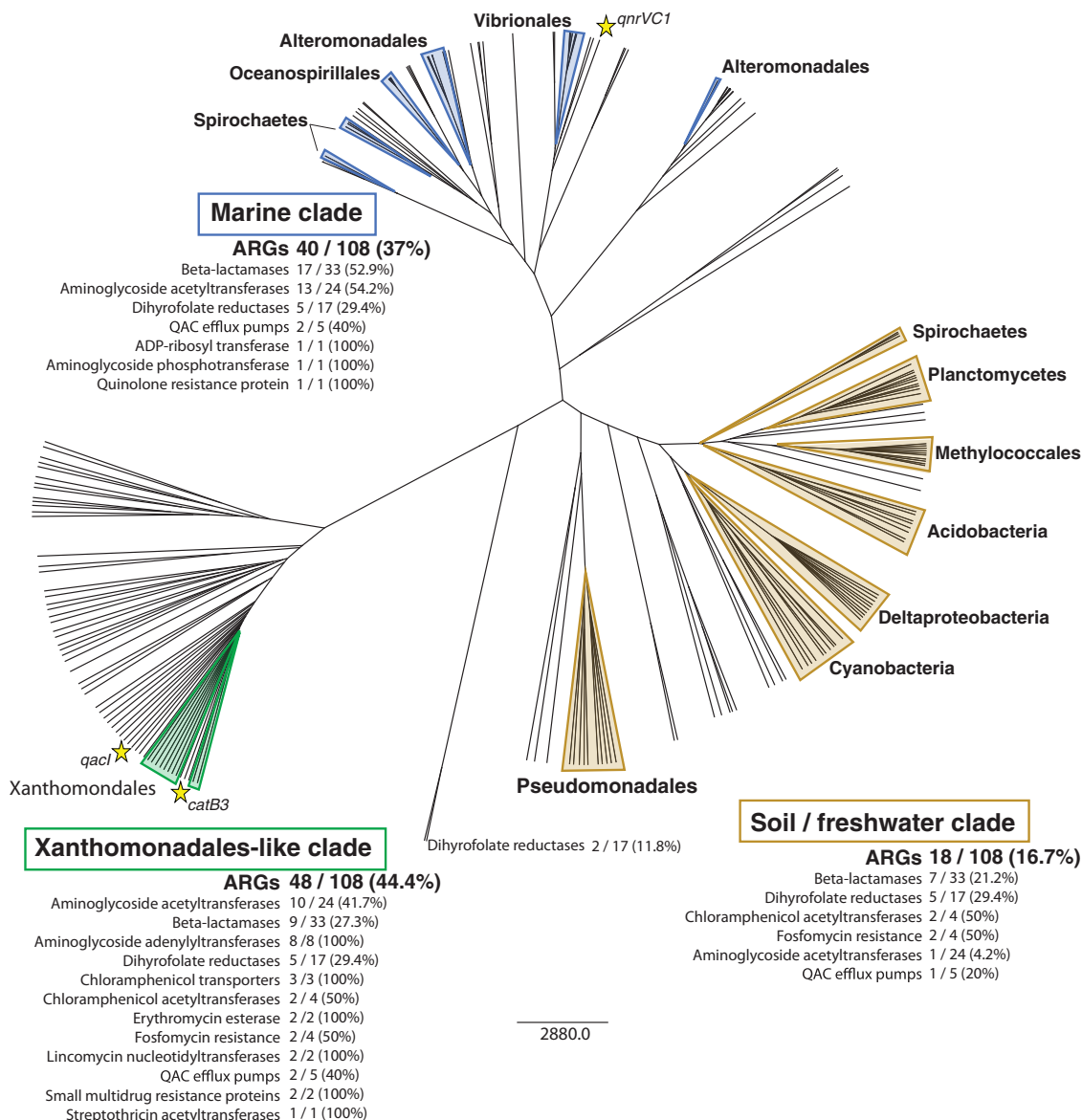


Fig. 2 Structure-based clustering of *attC*s. Ten top-scoring (based on covariance model (CM) bit scores) *attC*s for each taxon-specific CM are outlined by shaded triangles. Each non-shaded branch represents an *attC* site from one of 108 different resistance gene cassettes annotated by Partridge et al.¹⁰. The structures of *attC*s cluster according to host environment, forming three major clades. For each clade, the abundance of each resistance type is displayed, along with its relative proportion among all cassettes of that resistance type. Branches with yellow stars represent *attC*s from the three gene cassettes observed residing in the sedentary chromosomal integrons of the ancestral taxa predicted by our covariance models. We predicted that two of these (*qacI* and *catB3*) originated in Xanthomonadales, and the third (*qnrVC1*) in Vibrionales. Relative distances show all three are highly similar to their predicted taxa, further validating these predictions. See Supplementary Figure 1 for a tree with all branches labelled.

chromosome. These findings support our *attC* taxonomy predictions. We also show that *attC*s, based on structure and sequence, cluster according to the host environment rather than host phylogeny. We use this clustering to predict the source environments of each resistance cassette. Together, our findings represent the first evidence of the taxonomic and environmental origins of resistance cassettes.

The search for and use of novel antibiotics will inevitably result in the evolution and spread of novel resistance genes among pathogenic bacteria. Mobile integrons, will most likely play a role in their dissemination. Knowledge of the origins of already problematic resistance genes can allow us to predict hotspots for the emergence of the next generation of resistance genes before they enter clinical settings, and establish mechanisms to prevent them from doing so.

Methods

Collecting *attC*s from genomic sequences. Our approach to detect *attC* recombination sites was based on methods developed by Pereira et al.³⁵. This first involved using HattCI v1.0b³⁶, which uses a generalised hidden Markov model to detect the nucleotide sequence of each core motif of an *attC* site (i.e. R' - spacer' - L' - loop' - L' - spacer' - R'³⁷ (Fig. 1C)). HattCI was implemented so that both strands were analysed in batches of 40 sequences [parameters: -b -s 40 -t 8]. All output sequences generated by HattCI were then screened for the conserved *attC* folding structure. The consensus structure was generated using a structural alignment of 231 manually curated *attC*s, largely from class 1 and 2 integron cassette arrays, available in Supplementary Material Section B of Pereira et al.³⁶. The structural alignments were generated using LocARNA v1.9.2.1³⁸⁻⁴⁰, which uses tools built on the Turner free energy model to simultaneously fold and align input sequences. The alignments of the *attC*s were anchored so that all complementary conserved motifs were forced to align (i.e. R', L', L', R' (Fig. 1C)) [parameters: mlocarna --anchor-constraints --stockholm --threads 8]. A covariance model (CM) was built from the structural alignment using the cmbuild and cmcalibrate tools in the Infernal v1.1.2 package⁴¹ with default parameters. The CM was used to screen the HattCI output for the correct folding structure necessary

for *attC* functionality using Infernal's *cmsearch* tool with an E-value threshold of 0.01. Putative *attCs* were subject to further filtering to remove singletons, with the retained set being those that were clustered (at least two *attCs*) with no more than 4 kb between each⁵, that being twice the size of the largest annotated gene cassette⁴².

To collect *attCs* from distinct bacterial taxa, we first applied the above approach to screen an initial batch of complete genomes known to carry integrons⁵. We excluded any sequences that represented plasmids, as these would consist of *attCs* in cassettes from mobile integrons, which could have been acquired from a diverse set of taxa. The initial batch consisted of 1,825 complete chromosomal sequences, representing 20 bacterial phyla. We collected *attCs* for each phylum separately, and redundancy was removed from each set of *attCs* by discarding identical sequences with CD-HIT v4.6^{43,44} [parameters: *cdhit-est -c 1.0 -aL 1.0*]. We split the *attCs* from Gammaproteobacteria into order-level groupings due to the extensive number of *attCs* recovered and sequenced genomes present in this group. For those taxa in which we detected less than 50 non-redundant *attCs*, we sought additional chromosomal sequences from the NCBI Assembly database [accessed December 2020]. Any taxon that still had less than 50 representative *attCs* were excluded from further analysis.

Creating taxon-specific *attC* covariance models (CMs). For each taxon-specific set of *attCs*, we created a CM based on a structural alignment using methods described above. We then tested the efficacy of the models in predicting the host taxonomy of provided *attC* sequences. To do this, we searched each taxon-specific CM against our complete set of *attCs* ($n = 2,352$), using the *cmsearch* tool in the Infernal v1.1.2 package [parameters: *--cpu 8 --notrunc --nohmm*]. The efficacy of each CM was determined by its sensitivity, that being its capacity to detect *attCs* from the particular taxon that it was built from (true positives), and its specificity, which is the ability to exclude *attCs* derived from other taxa (false positives). CMs that did not achieve a specificity greater than 98% were excluded. All CMs are available as Supplementary Data 1.

Assigning ancestral host taxonomy to antibiotic resistance gene cassettes.

CMs that passed the efficacy screening were used to assign host taxonomy to a set of 108 resistance gene cassettes from mobile integrons, obtained from Table 1 in Partridge et al.¹⁰. The CMs were searched against the *attCs* from each resistance cassette using Infernal's *cmsearch* as described above. The bit score cut-offs for the taxonomic assignment were set for each CM individually in order to maximise sensitivity while ensuring specificity was 98–100% (See Supplementary Table 1 for specific bit scores).

In addition, we sought to determine if any of the resistance cassettes of the mobile integrons could be observed in their putative ancestral chromosomal context. ORFs of the resistance cassettes that could be assigned a taxonomy were aligned against all 24,143 complete bacterial chromosomes available in NCBI [downloaded on 27 January 2021]. Multiple alignments were implemented using BLAST v2.7.1⁴⁵ with 98% identity and 100% query cover thresholds [parameters: *-task megablast -perc_identity 98 -qcov_hsp_perc 100 -num_threads 8*]. Each hit was manually checked to exclude all instances of mobile integrons that had transposed into chromosomes as these are unlikely to be the ancestral sources.

We applied a structural-based clustering approach to visualise how similar *attCs* of the resistance cassettes were with representative *attCs* from each taxon. For this, we used RNAclust v1.3⁴⁶, which builds on locARNA to create a hierarchical-clustering tree from a WPGMA analysis. We collected the top 10 representative *attCs* for each CM based on their bit scores. These, along with the resistance cassette *attCs*, were clustered using RNAclust's default parameters.

Phylogenetic analysis of integron integrases. Integron integrases (IntIs) were obtained from selected genomes that had the top-scoring *attCs* for each CM. We aimed to collect five representative IntI sequences from each taxon using a profile HMM provided by Cury et al.⁵. The profile HMM was based on the additional domain that is unique to integron integrases, separating them from other tyrosine recombinases⁴⁷. For each selected genome, proteins were annotated using Prodigal v2.6.3⁴⁸ and the profile HMM was used to determine which sequences represented IntIs with the *hmmsearch* tool from the HMMER v3.2 package⁴⁹. Any partial IntI sequences were discarded from the phylogenetic analysis and the remainder were manually screened to ensure that they possessed the complete additional domain. The orientation of the *intI* gene in relation to the cassette array was confirmed using IntegronFinder v1.5.1⁵ [parameters: *--local_max --func_anot*].

IntI sequences were aligned using MAFFT v7.271⁵⁰ [parameters: *--localpair --maxiterate 1000*]. The best substitution model to suit the alignment was determined using ModelFinder⁵¹, which was used to construct a maximum-likelihood tree with IQ-TREE v1.6.12^{52,53} with 1,000 bootstrap replicates [parameters: *-m MFP -alrt 1000 -bb 1000 -nt 8*].

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

Data availability

Specific bacterial genomes were downloaded from NCBI via accessions provided by Cury et al.⁵ using the following Perl command [*perl -e 'use LWP::Simple;getstore("http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&rettype=fasta&retmode=text&id=" . join(" ", qw(<space-separated list of nucleotide accession numbers>)) . ",output_filename=" . fasta";]*]. The remaining genome sequences were downloaded directly from the NCBI Assembly database (<https://www.ncbi.nlm.nih.gov/assembly>).

Code availability
Software used in this study are LocARNA v1.9.2.1; HattCI v1.0b; Infernal v1.1.2; CD-HIT v4.6; BLAST v2.7.1; RNAclust v1.3; Prodigal v2.6.3; HMMER v3.2; IntegronFinder v1.5.1; MAFFT v7.271; IQ-TREE v1.6.12; ViennaRNA v2.0. Specific parameters used for each software are provided in detail in the Methods section.

Received: 12 April 2021; Accepted: 16 July 2021;

Published online: 09 August 2021

References

- Gillings, M. et al. The evolution of class 1 integrons and the rise of antibiotic resistance. *J. Bacteriol.* **190**, 5095–5100 (2008).
- Ghaly, T. M., Geoghegan, J. L., Alroy, J. & Gillings, M. R. High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environ. Microbiol.* **21**, 1567–1574 (2019).
- Gillings, M. R. Integrons: past, present, and future. *Microbiol. Mol. Biol. Rev.* **78**, 257–277 (2014).
- Mazel, D. Integrons: agents of bacterial evolution. *Nat. Rev. Microbiol.* **4**, 608–620 (2006).
- Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
- Cambray, G. et al. Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mob. DNA* **2**, 6 (2011).
- Guerin, É. et al. The SOS response controls integron recombination. *Science* **324**, 1034–1034 (2009).
- Escudero, J. A., Loot, C., Nivina, A. & Mazel, D. The integron: adaptation on demand. *Microbiol. Spectr.* **3**, MDNA3-0019–MDNA3-0012014 (2015).
- Ghaly, T. M., Chow, L., Asher, A. J., Waldron, L. S. & Gillings, M. R. Evolution of class 1 integrons: mobilization and dispersal via food-borne bacteria. *PLoS One* **12**, e0179169 (2017).
- Partridge, S. R., Tsafnat, G., Coiera, E. & Iredell, J. R. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol. Rev.* **33**, 757–784 (2009).
- Gillings, M. R. Class 1 integrons as invasive species. *Curr. Opin. Microbiol.* **38**, 10–15 (2017).
- Zhu, Y.-G. et al. Microbial mass movements. *Science* **357**, 1099–1100 (2017).
- Rowe-Magnus, D. A. et al. The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc. Natl Acad. Sci.* **98**, 652–657 (2001).
- Cambray, G., Guerout, A.-M. & Mazel, D. Integrons. *Annu. Rev. Genet.* **44**, 141–166 (2010).
- Bouvier, M., Demarre, G. & Mazel, D. Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J.* **24**, 4356–4367 (2005).
- Mukhortava, A. et al. Structural heterogeneity of *attC* integron recombination sites revealed by optical tweezers. *Nucleic Acids Res.* **47**, 1861–1870 (2018).
- Loot, C., Ducos-Galand, M., Escudero, J. A., Bouvier, M. & Mazel, D. Replicative resolution of integron cassette insertion. *Nucleic Acids Res.* **40**, 8361–8370 (2012).
- Nivina, A. et al. Structure-specific DNA recombination sites: design, validation, and machine learning-based refinement. *Sci. Adv.* **6**, eaay2922 (2020).
- Nivina, A., Escudero, J. A., 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.* **44**, 7792–7803 (2016).
- Bouvier, M., Ducos-Galand, M., Loot, C., Bikard, D. & Mazel, D. Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genet.* **5**, e1000632 (2009).
- Biskri, L., Bouvier, M., Guérout, A.-M., Boissard, S. & Mazel, D. Comparative study of class 1 integron and *Vibrio cholerae* superintegron integrase activities. *J. Bacteriol.* **187**, 1740–1750 (2005).
- Larouche, A. & Roy, P. H. Effect of *attC* structure on cassette excision by integron integrases. *Mob. DNA* **2**, 3 (2011).
- Gillings, M. R., Holley, M. P., Stokes, H. W. & Holmes, A. J. Integrons in *Xanthomonas*: a source of species genome diversity. *Proc. Natl Acad. Sci. U. S. A.* **102**, 4419–4424 (2005).

24. Rowe-Magnus, D. A., Guerout, A.-M., Biskri, L., Bouige, P. & Mazel, D. Comparative analysis of superintegrations: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res.* **13**, 428–442 (2003).
25. Vaisvila, R., Morgan, R. D., Posfai, J. & Raleigh, E. A. Discovery and distribution of super-integrations among Pseudomonads. *Mol. Microbiol.* **42**, 587–601 (2001).
26. Ebmeyer, S., Kristiansson, E. & Larsson, D. G. J. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Commun. Biol.* **4**, 8 (2021).
27. Ghaly, T. M., Geoghegan, J. L., Tetu, S. G. & Gillings, M. R. The peril and promise of integrons: Beyond antibiotic resistance. *Trends Microbiol.* **28**, 455–464 (2020).
28. Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605–608 (1998).
29. Boucher, Y., Labbate, M., Koenig, J. E. & Stokes, H. W. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.* **15**, 301–309 (2007).
30. Timilsina, S. et al. Xanthomonas diversity, virulence and plant–pathogen interactions. *Nat. Rev. Microbiol.* **18**, 415–427 (2020).
31. Hayward, A. C., Fegan, N., Fegan, M. & Stirling, G. R. *Stenotrophomonas* and *Lysobacter*: ubiquitous plant-associated gamma-proteobacteria of developing significance in applied microbiology. *J. Appl. Microbiol.* **108**, 756–770 (2010).
32. Fitzpatrick, C. R. et al. Assembly and ecological function of the root microbiome across angiosperm plant species. *Proc. Natl Acad. Sci.* **115**, E1157–E1165 (2018).
33. Gu, Y. et al. The effect of microbial inoculant origin on the rhizosphere bacterial community composition and plant growth-promotion. *Plant Soil* **452**, 105–117 (2020).
34. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
35. Buongiorno Pereira, M. et al. A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* **21**, 495 (2020).
36. Pereira, M. B., Wallroth, M., Kristiansson, E. & Axelson-Fisk, M. HattCI: Fast and accurate *attC* site identification using hidden Markov models. *J. Comput. Biol.* **23**, 891–902 (2016).
37. Stokes, H., O’gorman, D., Recchia, G. D., Parsekhian, M. & Hall, R. M. Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol. Microbiol.* **26**, 731–745 (1997).
38. Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biol.* **3**, e65 (2007).
39. Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012).
40. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
41. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
42. Joss, M. J. et al. ACID: annotation of cassette and integron data. *BMC Bioinforma.* **10**, 118 (2009).
43. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
44. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
45. Madden, T. in *The NCBI Handbook [Internet]. 2nd edition* (eds et al.) (National Center for Biotechnology Information (US). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK153387/>. 2013).
46. Engelhardt, J., Heyne, S., Will, S. & Reiche, K. *RNAclust Documentation*, <http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAclust/manual.pdf> (2010).
47. Messier, N. & Roy, P. H. Integron integrases possess a unique additional domain necessary for activity. *J. Bacteriol.* **183**, 6699–6706 (2001).
48. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
49. Eddy, S. R. HMMER 3.2: Biosequence analysis using profile hidden Markov models. <http://hmmer.org/> (2018).
50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evolution* **30**, 772–780 (2013).
51. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
52. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evolution* **35**, 518–522 (2018).
53. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evolution* **32**, 268–274 (2015).

Acknowledgements

This research was supported by the Australian Research Council Discovery Grant DP200101874. TMG would like to thank Mary and Saoirse Ghaly for loving support.

Author contributions

TMG contributed to the conception of the study, performed all data analyses, wrote the original draft of the paper, and contributed to the final editing and revision of the paper. SGT contributed to the conception of the study and the final editing and revision of the paper. MRG contributed to the conception of the study and the final editing and revision of the paper. All authors contributed to the article and approved the final submitted version.

Competing interests

The authors declare no competing interests.

Additional information


Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02489-0>.

Correspondence and requests for materials should be addressed to T.M.G.

Peer review information *Communications Biology* thanks Johan Bengtsson-Palme and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Audrone Lapinaite and Luke R. Grinham. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021