MACQUARIE University
SYDNEY·AUSTRALIA

Ecology of the integron gene cassette metagenome

Timothy M. Ghaly

Supervised by Prof. Michael Gillings and Dr Jemma Geoghegan

Department of Biological Sciences

Macquarie University, New South Wales, 2109

Submitted:

10 September 2018 as part of the requirements for completion of the degree of

Master of Research

Table of contents

Declaration

This work is presented as a 'thesis by publication'. Chapter 2 is written as a manuscript for submission to *Proceedings of the National Academy of Sciences* and follows the journal's guidelines set out for publication.

I declare that the work in this thesis has not been submitted for a higher degree to any other university or institution.

I wish to acknowledge the Environmental and Molecular Microbiological Analysis Laboratory, Macquarie University, for providing environmental DNA samples from soils sampled from Herring Island, Antarctica and Sturt National Park, New South Wales, Australia.

All other research described in this thesis is my own original work.

Timothy M. Ghaly
10 September 2018

Acknowledgements

I would like to thank both my supervisors, Michael Gillings and Jemma Geoghegan, for their advice and guidance throughout my project. Many thanks go to my fellow colleagues in the EMMA Lab. I would also like to thank John Alroy for his assistance in obtaining ecological richness estimates, and Sasha Tetu and Martin Ostrowski for advice on bioinformatic analyses.

A special thanks is owed to my partner Mary Nolan for love and support as well as for comments on earlier drafts of the manuscript.

Summary

Integrons are genetic elements that promote rapid adaptation in bacteria by capturing exogenous mobile gene cassettes. Recently, a sub-set of gene cassettes have facilitated the global spread of antibiotic resistance, however, outside of clinical settings, very little is known about the function and ecology of these cassettes. Here, I sequenced whole cassettes from soils sampled across Australia and Antarctica, and recovered 44,970 cassettes that encoded 27,215 unique proteins. This represents an order of magnitude more cassettes than previous sequencing efforts. Cassettes had extremely high local richness, with estimates ranging from 4,000 to 18,000 unique cassettes per 0.3 grams of soil. Gene cassettes exhibited a rapid spatial turnover and had a heterogeneous distribution across space. More than 84% encoded unknown proteins, 64% of which had no homologs in existing databases. These findings provide insights into gene cassette ecology, and highlight the diversity in this metagenome. This diversity can generate genomic complexity and drive bacterial evolution. I also explore the potential use of integron gene cassettes in accelerating the discovery of novel proteins. The gene cassette metagenome represents a huge untapped resource that provides an efficient means to shed light on the dark matter of the protein universe. This resource is thus of substantial biotechnological importance, particularly for developing small-molecule therapeutics and engineering molecular tools.

Chapter 1 – Introduction

Types of integrons

Integrons are genetic elements that promote rapid adaptation in bacteria by capturing, rearranging, and expressing mobile gene cassettes. These elements are common in diverse bacterial genomes and they have access to a vast pool of gene cassettes, known as the cassette metagenome. Integrons themselves are extremely diverse, being categorised into hundreds of classes (1-3), which together with their broad taxonomic distribution, suggest they are ancient elements.

In recent years, one type of integron, the class 1 integron, has received much attention because of its role in the dissemination of antibiotic resistance determinants (5). Class 1 integrons, due to their association with mobile DNA elements, particularly the Tn*402* family of transposons (6), have transferred into at least 72 bacterial species (7), and have collectively acquired over 130 different antibiotic resistance genes (8).

All integrons of clinical relevance (classes 1 to 5) are found embedded in transposons and conjugative elements, providing a distinction with chromosomal integrons, that lack this mobility. In general, mobile integrons carry few cassettes, generally encoding resistance to antibiotics. Chromosomal integrons encompass hundreds of integron classes and can carry a much larger number of gene cassettes of mostly unknown function. Some *Vibrio* species are known to carry hundreds of cassettes within a single integron (9, 10). These large chromosomal integrons have been referred to as 'super-integrons'. Chromosomal integrons are believed to be the ancestors of mobile integrons (11, 12).

Integrons have likely been present in bacterial genomes for hundreds of millions of years. This is evident from the phylogeny of chromosomal integrons, which is roughly congruent with the 16S rDNA phylogeny of their hosts (11, 13). Thus, the occurrence of chromosomal integrons likely predates host speciation.

The distinction between mobile and super-integrons is not as clear cut as has been portrayed in the literature, since there is a continuous spectrum of integron structures (14, 15), nevertheless these terms highlight some important characteristics. Mobile integrons facilitate rapid lateral transfer into a broad range of taxa, along with the genes they carry, while chromosomal integrons drive genomic complexity and phenotypic diversity within lineages (16).

The integron platform: structure and function

All integrons have three key features that together ensure the successful insertion and expression of exogenous gene cassettes. The first is *intI*, a gene that encodes an integron integrase (IntI), responsible for catalysing the insertion of gene cassettes at the second core feature, the integron

recombination site (*attI*) (Fig. 1a) (17). The third feature is the integron promoter (Pc), which drives the transcription of integron gene cassettes, with decreasing expression with distance from the promoter (Fig. 1a) (18).

In general, gene cassettes are promoterless, circular open reading frames (ORFs) that possess a cassette-associated recombination site (*attC*) (19). Gene cassettes are commonly integrated via recombination between the *attC* site of the cassette, and the *attI* site of the integron platform (Fig. 1a). Multiple gene cassettes can be sequentially inserted to form an integron cassette array. Cassettes may be lost over time, as any gene cassette can be excised from the array. This occurs through recombination between two *attC* sites, resulting in the excision of the intermediate gene cassette as a circular DNA molecule (Fig. 1b). In some cases, the excised circular gene cassette may be reinserted at the *attI* site, bringing the cassette closer to the integron promoter where its expression can be maximised (Fig. 1b).

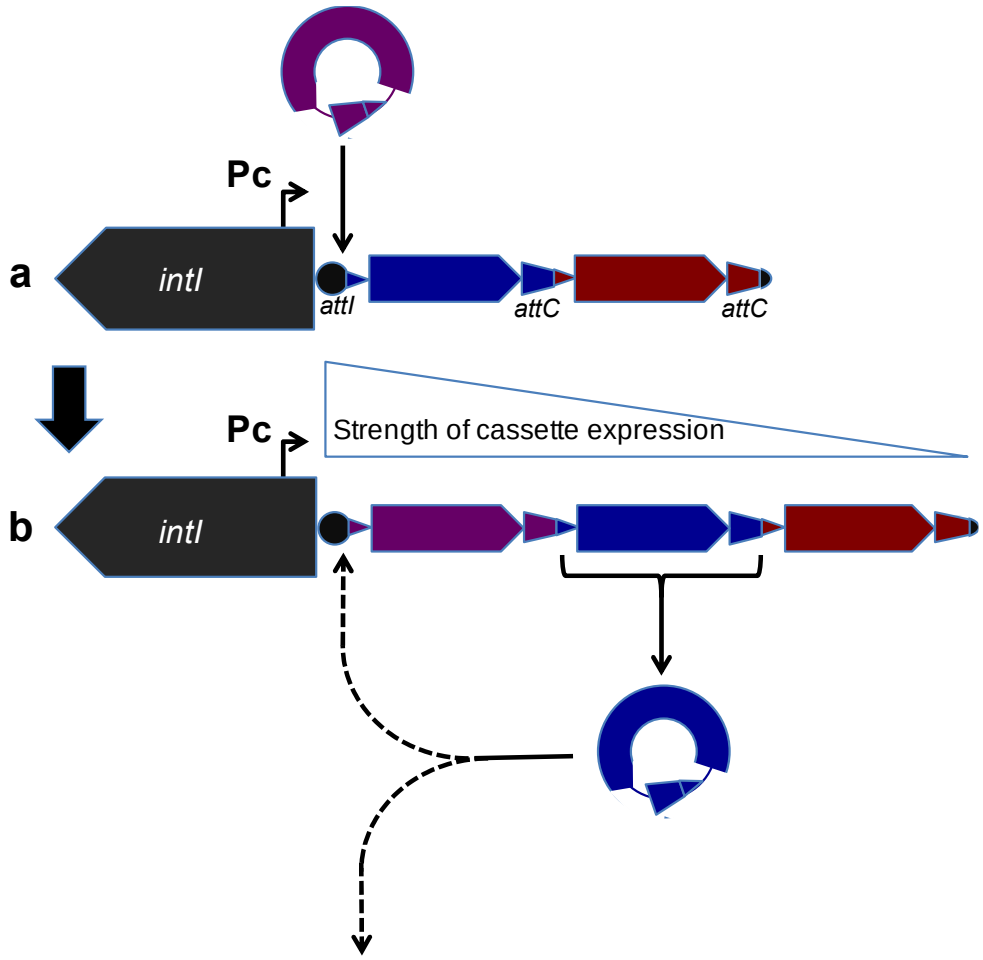

Fig. 1. Integron structure and model for gene acquisition and excision. All integrons carry an integron integrase gene (*intI*), an integron recombination site (*attI*), and gene cassette promoter (Pc). (**a**) Incoming gene cassettes are inserted into the integron at the *attI* site via *attC* x *attI* recombination. (**b**) Any gene cassette can be excised from the cassette array via *attC* x *attC* recombination and may be lost or reinserted at the *attI* site where expression is maximised.

The insertion and excision of gene cassettes relies on *attC* sites, which act as recombination substrates recognised by IntI. The structure of *attC* sites involves two core features: the R"-L" arm and the L'-R' arm, separated by a highly variable spacer (16-109 bp) (Fig. 2a) (20, 21). Among these domains, only R" and R' have conserved sequences: 5'-RYYYAAC and 5'-GTTRRRY, respectively. Although there is a lack of sequence conservation between *attC* sites, they do exhibit a highly conserved palindromic arrangement. The palindromic arrangement allows the formation of a single-stranded cruciform structure by pairing R" with R' and L" with L' (Fig. 2b) (22, 23).

Since gene cassettes are promoterless, it is essential that they be inserted at the *attI* site in the correct orientation, to allow expression from the integron promoter. This is ensured by IntI binding specifically to the cruciform of the bottom *attC* strand ($attC_{bs}$) only, which is recognised by the presence of two to three extrahelical bases (Fig. 2b) (2, 24-27). Therefore cassette insertion (*attI* X *attC* recombination) involves only the bottom strand of the *attC* site, forming a single-stranded Holliday junction, which is resolved after replication (28). Since *attC* sites rely more on their structure than their sequence, evolutionarily distinct IntI proteins can capture any available gene cassette, which collectively carry very different *attC* sequences. This makes integrons extremely successful loci for generating rapid diversity and novel adaptive phenotypes.

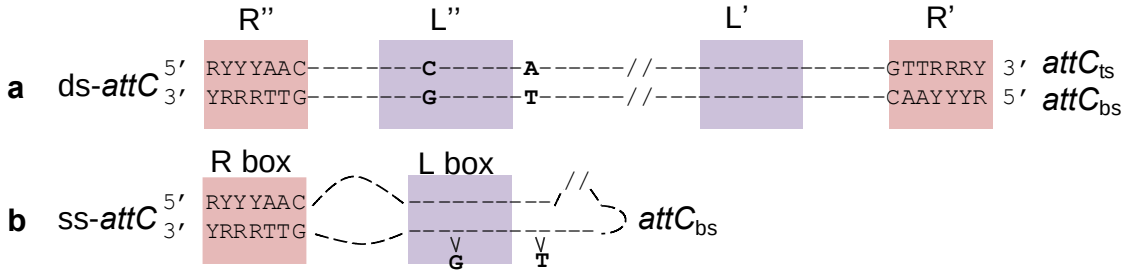

Fig. 2. Conserved structure of gene cassette *attC* sites. (**a**) Palindromic structure of the non-recombinogenic double-stranded *attC* (ds-*attC*). (**b**) The folded cruciform structure of the recombinogenic bottom *attC* strand ($attC_{bs}$). Extrahelical bases that extrude out of the folded single-stranded structure are shown in bold.

Evolutionary outcomes of gene cassette insertion and excision

Cassette insertion involves the formation of a single-stranded *attC* structure, which is then resolved via replication (28). Cassette insertion is mediated by IntI, a member of the tyrosine recombinase family, which also includes the λ phage integrase and the P1 phage Cre recombinase. In general, this family of enzymes follows a common pathway of site-specific recombination as follows. First, the recombinase cleaves a single strand of DNA from each of the recombining substrates (29). This allows the first strand exchange between the two substrates, forming a recombination structure

known as a Holliday junction (Fig. 3a) (29). This structure is then resolved by a second strand exchange, completing the recombination reaction (Fig. 3a) (29).

Cassette insertion, however, involves the recombination between a double-stranded *attI* site and a single-stranded *attC* site (bottom strand only), thus ensuring correct orientation of the inserted cassette (25, 26). The resulting recombination structure represents an asymmetrical Holliday junction (Fig. 3b). Consequently, resolution of this structure via the typical second strand exchange would result in abortive recombination products. Instead, this atypical Holliday junction is resolved after replication of the complete molecule (28). Therefore, an incoming gene cassette will only be inserted into one of the two daughter molecules. Replication of the recombinogenic strand resolves the Holliday junction, resulting in the integration of the gene cassette, while replication of the alternate strand generates the original substrate prior to the recombination event (Fig. 3b).

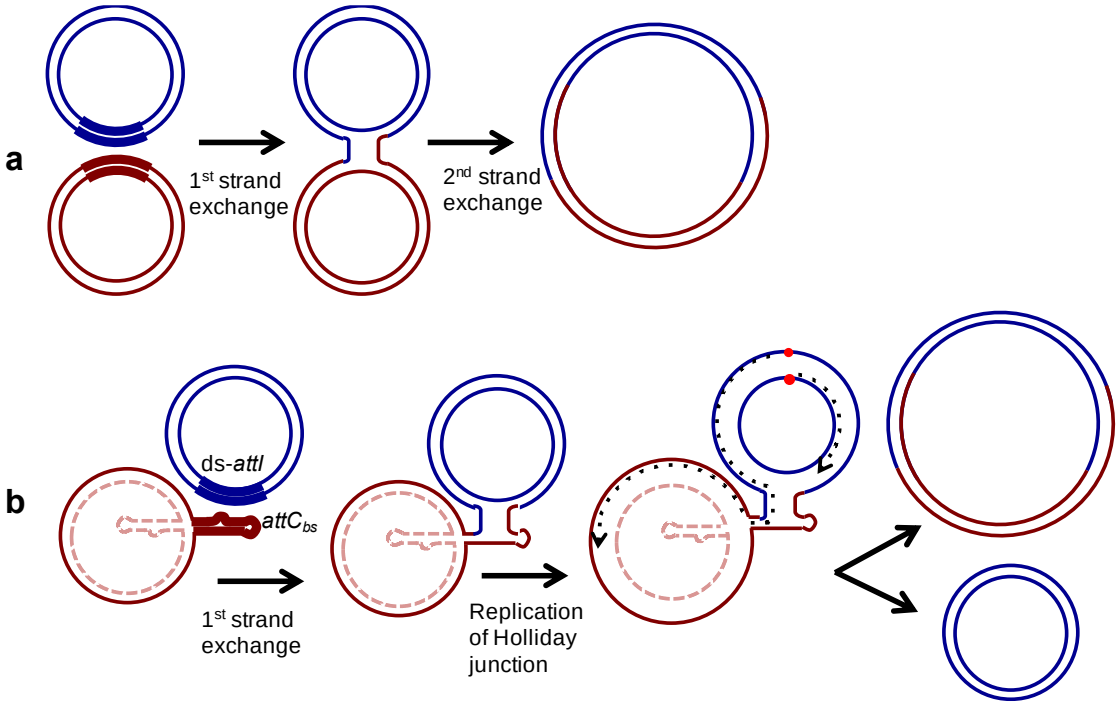

Fig. 3. Two different pathways for site-specific recombination. (**a**) The general recombination pathway catalysed by members of the tyrosine recombinase family. A single strand on each of the recombining substrates is cleaved, allowing for a single strand exchange. This generates a symmetrical Holliday junction, which is resolved after a second strand exchange. (**b**) The integron recombination pathway during cassette insertion involves a single strand exchange between the double stranded *attI* site (ds-*attI*) and the bottom stand of the *attC* site (*attC*$_{bs}$). This generates an atypical Holliday junction, which is resolved after replication (dotted black arrows; lagging strand not shown) of the complete molecule. Replication results in one recombinant daughter molecule that possesses the inserted cassette, and one daughter molecule that represents the original substrate prior to the recombination reaction.

Replicative resolution thus generates cassette array diversity between daughter molecules. In turn, this allows for direct competition between the two daughter integrons, one with the newly inserted cassette, and one without. Whichever daughter molecule provides the greatest selective advantage will outcompete the other and be driven to fixation. This provides the potential for any

available gene cassette to be sampled with the possibility of returning to the original genotype if the novel cassette reduces host fitness.

Since replicative resolution during cassette insertion (*attI* x *attC* recombination) is a product of the single-stranded folding of *attC*, the same mechanism is likely to be involved in cassette excision (*attC* x *attC* recombination) (28). This implies that the excision and reinsertion of a cassette can result in gene duplication. The excision of a gene cassette will generate one daughter molecule that still carries the excised cassette after replication. Consequently, the reinsertion of this cassette at the *attI* site can give rise to an integron that carries two copies of this gene (28). Gene duplication has long been recognised as a driving force in evolution (30, 31). It provides the raw material for the evolution of new gene functions, where mutations can accumulate in one of the gene copies. These mutations can lead to the diversification of gene cassettes and the generation of altered or novel functions. Indeed, we see multiple families of gene cassettes that share high sequence homology spanning the whole ORF and *attC* site (8, 32). This diversity may be, at least in some part, a product of gene duplication events. Replicative resolution may help explain not only the diversification of integron cassette arrays, but also of the gene cassettes themselves.

Gene cassette genesis

It is still unclear in which genomic backgrounds gene cassettes originate and how they are initially formed. The fact that gene cassettes are generally comprised of a single ORF with minimal non-coding sequence, and lack a promoter region, has led to the hypothesis that cassettes are generated by reverse transcription of an mRNA molecule (19, 33). Under this hypothesis, the *attC* site may be added to the gene, either prior to or after the reverse transcription event, to result in the final structure of an integron gene cassette.

Since *attC* sites resemble rho-independent transcriptional terminators, they may have been present in the original transcripts. This in turn would provide a suitable priming site for some as yet unidentified reverse transcriptase. This *attC*-primer RT model, however, has been disputed (34, 35). This is largely because a number of gene cassettes, which have their own promoter, or are in the reverse orientation relative to their *attC* site, do not fit this model. A noteworthy example is that of the toxin-antitoxin (TA) gene cassettes, which possess their own promoter required for autoregulation. TA systems are selfish genes that ensure their own propagation by causing post-segregational death of any daughter cell that does not inherit the TA locus (36). TA cassettes, which are often found in large super-integrons, cause the death of any cell in which the TA cassettes are lost. All TA systems are autoregulated by the binding of the antitoxin to their own promoter (37). Thus, the presence of a promoter within TA cassettes is inconsistent with this model, since the promoter motif cannot be present in the original RNA transcript.

The reverse transcription hypothesis has since been updated and proposes gene cassette formation to be mediated by group II introns. Bacterial group II introns are both ribozymes and mobile elements. They are comprised of a catalytic RNA and an intron-encoded protein, containing reverse transcriptase activity. The intron-encoded protein mediates the self-splicing of its mRNA transcript, to which it subsequently binds, forming a ribonucleoprotein complex. The intron RNA is then inserted into a DNA target site, where it is reverse-transcribed by the intron-encoded protein (38). In particular, group IIC introns are known to insert immediately before or after transcriptional terminators or similar stem-loop motifs (39, 40). Interestingly, several cases of IIC introns inserted immediately prior to *attC* sites have now been observed (41-44). The intron-encoded proteins of these IIC-*attC* introns have been shown, at least *in vitro*, to exhibit both RNA-dependent and DNA-dependent DNA polymerase activity (4). This has led to the proposal of a specific gene cassette formation model via IIC introns (Fig. 4) (4, 42).

The proposed model involves two independent intron insertion events (Fig. 4a). The result is one intron downstream of a gene, inserting prior to its transcriptional terminator, and another intron copy in an isolated *attC* site (Fig. 4b). Homologous recombination between the two intron copies would bring the gene and the *attC* site together, forming a gene-intron-*attC* intermediate (Fig. 4c). Self-splicing of the intron from the transcribed intermediate (Fig. 4d-e), followed by reverse transcription of the gene-*attC* RNA template would lead to the formation of a DNA gene cassette (Fig. 4f) (4). The newly formed cassette could then be inserted into an integron platform (Fig. 4g).

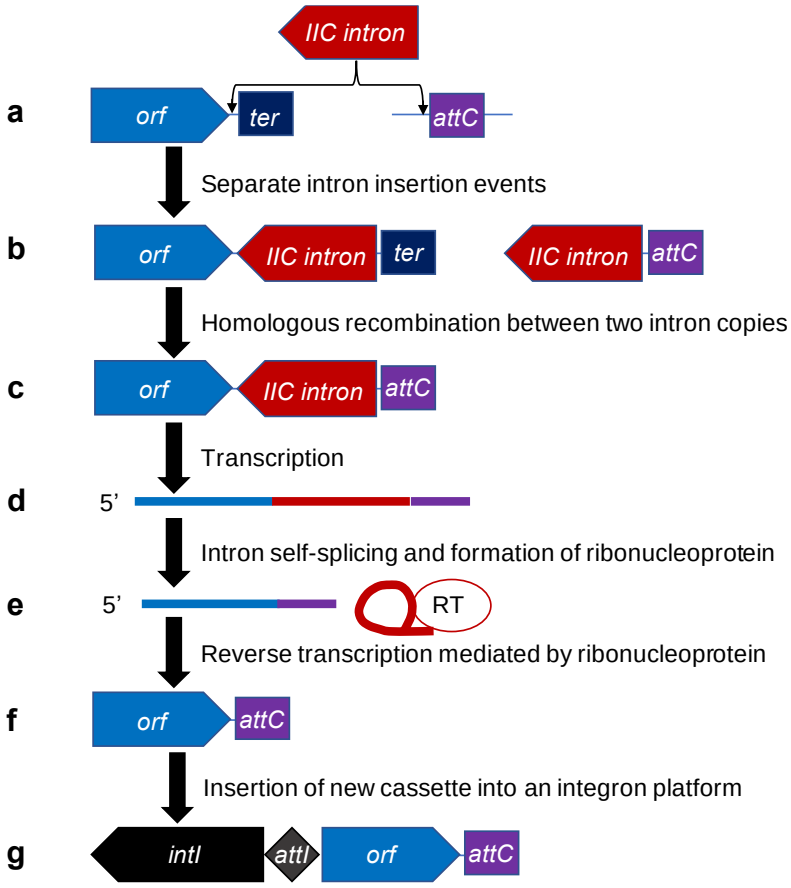

Fig. 4. Proposed model for gene cassette formation via group IIC introns (4). (a) Two separate intron insertion events, resulting in (b) one intron copy immediately downstream from a gene (*orf*), prior to its transcriptional terminator (*ter*), and a second copy prior to an isolated *attC* site. (c) Homologous recombination between the two intron copies results in an *orf*-intron-*attC* fusion product. (d) The *orf*-intron-*attC* product is transcribed into mRNA. (e) The intron RNA is spliced out of the transcript, forming an RNA lariat that associates with the intron-encoded reverse transcriptase protein (RT) to form a ribonucleoprotein complex. (f) Reverse transcription of the mRNA molecule, mediated by the ribonucleoprotein complex, generates a DNA gene cassette, which can be (g) inserted into an integron platform at an integron-associated attachment site (*attI*).

This model may also address the issues concerning the traditional *attC*-primer RT hypothesis. Gene cassettes that carry an ORF in the reverse orientation could be explained by the initial intron insertion site occurring after the transcriptional terminator, rather than before it. This would instead allow for the capture of the downstream ORF, which may run in the reverse orientation. Such instances would generally prevent expression of the cassette once inserted into an integron, and thus be rarely observed. Furthermore, the synthesis of TA cassettes carrying promoters could be explained by the possible DNA-dependent DNA polymerase activity of IIC introns, although an exact pathway is yet to be presented.

Gene cassette diversity and functionality

Gene cassettes associated with mobile integrons largely confer resistance to antimicrobial agents. More than 130 different antibiotic resistance genes, collectively conferring resistance to nearly all classes of antibiotics, have been found associated with mobile integrons (8).

Chromosomal gene cassettes, however, encode for proteins of largely unknown functions. An analysis of 1,677 gene cassettes from *Vibrionales* revealed 67% of cassettes encoded completely novel proteins, while a further 13% encoded proteins with homologs of unknown functions (3). Similarly, Koenig et al. (45) sequenced 2,145 cassettes from four marine sediment samples, and found ~80% encoded proteins with no known homologs (45). There is a small proportion of environmental cassettes with characterised homologs. These encode proteins with a wide range of predicted functions. The most prevalent of these include acetyltransferases, DNA modification, phage-related functions, and TA systems (2, 3). Moreover, 10% to 30% of cassette-encoded proteins contain signal peptide domains necessary for membrane association or cellular export (45, 46), and about 30% of cassettes encode proteins with transmembrane domains (46). The very few cassette-encoded proteins that have had their functions experimentally determined include restriction or methylation systems, sulfate-binding proteins, lipases, polysaccharide biosynthesis, and dNTP pyrophosphohydrolases (11, 47, 48). Together, these findings suggest gene cassettes can provide significant adaptive potential to bacteria and facilitate their interactions with the surrounding environment.

Analysing the protein-encoded diversity of the gene cassette metagenome could accelerate contributions to protein discovery. Existing methods, specifically whole genome and environmental metagenomic sequencing, often fail to capture the vast pool of novel proteins that are encoded by rare genes. Such rare gene clusters are unlikely to be represented within environmental DNA libraries (49). This is demonstrated by the large proportion of proteins that can be assigned functions using such methods. Specific functions can be assigned to 76% of ORFs predicted from environmental metagenomic data, and 83% of ORFs for completely sequenced genomes (50). These proportions increase to 83% and 86%, respectively, when non-specific functions are considered (50).

To overcome this limitation in protein discovery, targeting of rare genes for sequencing may be used to prospect the unknown regions of the protein universe. PCR amplification has been successfully used to recover entire gene cassettes from environmental DNA (1, 45, 51-53). Conserved regions within the *attC* site of cassettes allow PCR priming (51). This could facilitate protein discovery at an unprecedented rate, given the high proportion of novel proteins encoded by integron gene cassettes. The gene cassette metagenome thus represents a useful tool that can help shed light on the dark matter of the protein universe. This is important for understanding the vast array of traits shared by bacteria, and is also of substantial biotechnological importance, particularly for developing small-molecule therapeutics and engineering molecular tools.

How big is the gene cassette metagenome?

Exploration of the cassette metagenome show that it is diverse and ubiquitous. Gene cassettes have been recovered from every environment surveyed, including soils and aquatic sediments, human and animal microbiomes, the phyllosphere, aquatic biofilms, seawater, and deep-sea hydrothermal vents (6, 54-59). These observations suggest that the cassette metagenome encompasses extensive genetic novelty, however, the extent of gene richness captured by the cassette metagenome is still unknown.

Several studies have attempted to address this question, but none has achieved the sequencing depth, nor sampled at the appropriate spatial scales required to gain an accurate richness estimate. Michael et al. (53) estimated a richness of 2,343 cassettes within a 50 m^2 sampling area. This estimate, however, was based on cassette length, determined by polyacrylamide gel electrophoresis, rather than sequence data. In this study, cassette 'types' were defined by size differences, with the assumption that there were only two cassettes per size class. As such, their richness calculation is likely to be a gross underestimate. The first sequence-based approach estimated a richness of ~3,000 gene cassettes for four marine sediment samples (45). This,

however, relied upon cloning of PCR products to obtain DNA sequences. Consequently, the depth of sequencing was unlikely to be sufficient to provide an accurate cassette richness estimate.

To overcome these issues, a shotgun sequencing approach may provide an appropriate depth of sampling. However, to obtain a meaningful richness estimate, sampling should also be implemented over large spatial scales. This is of particular importance, given that soil samples taken from 1-m intervals reveal different PCR-amplification profiles (51), suggesting significant spatial heterogeneity of gene cassettes. Determining the rate of spatial turnover and distribution of gene cassettes is the first step in understanding the true diversity within the gene cassette metagenome.

In this thesis, I have examined the spatial distribution of integron gene cassettes across sites from Australia and Antarctica to assess the biogeography of their functional gene diversity. I found that gene cassettes exhibit extremely high local richness, largely encoded novel proteins, have a rapid turnover through space, and show great spatial heterogeneity. The gene cassette metagenome is thus a vast pool of genetic diversity that can facilitate bacterial genome evolution, and can shed light on the dark matter of the protein universe.

References

1. Abella J, Fahy A, Duran R, & Cagnon C (2015) Integron diversity in bacterial communities of freshwater sediments at different contamination levels. *FEMS Microbiol Ecol* 91(12):fiv140.
2. Cambray G, Guerout A-M, & Mazel D (2010) Integrons. *Annu Rev Genet* 44:141-166.
3. Boucher Y, Labbate M, Koenig JE, & Stokes H (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol* 15(7):301-309.
4. Léon G & Roy PH (2009) Potential role of group IIC-*attC* introns in integron cassette formation. *J Bacteriol* 191(19):6040-6051.
5. Gillings MR (2017) Class 1 integrons as invasive species. *Curr Opin Microbiol* 38:10-15.
6. Ghaly TM, Chow L, Asher AJ, Waldron LS, & Gillings MR (2017) Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS One* 12(6):e0179169.
7. Domingues S, da Silva GJ, & Nielsen KM (2015) Global dissemination patterns of common gene cassette arrays in class 1 integrons. *Microbiology* 161(7):1313-1337.
8. Partridge SR, Tsafnat G, Coiera E, & Iredell JR (2009) Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 33(4):757-784.
9. Chen C-Y, *et al.* (2003) Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res* 13(12):2577-2587.
10. Escudero JA, Loot C, Nivina A, & Mazel D (2015) The integron: adaptation on demand. *Mobile DNA III*, eds Craig N, Chandler M, Gellert M, Lambowitz A, Rice P, & Sandmeyer S (American Society of Microbiology Press, Washington, DC.), pp 139-161.
11. Rowe-Magnus DA, *et al.* (2001) The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc Natl Acad Sci U S A* 98(2):652-657.
12. Gillings M, *et al.* (2008) The evolution of class 1 integrons and the rise of antibiotic resistance. *J Bacteriol* 190(14):5095-5100.
13. Rowe-Magnus DA & Mazel D (2001) Integrons: natural tools for bacterial genome evolution. *Curr Opin Microbiol* 4(5):565-569.
14. Hall RM, Holmes AJ, Roy PH, & Stokes H (2007) What are superintegrons? *Nat Rev Microbiol* 5(2):162.
15. Hall RM & Stokes H (2004) Integrons or super integrons? *Microbiology* 150(1):3-4.
16. Boucher Y, *et al.* (2011) Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *MBio* 2(2):e00335-00310.
17. Partridge SR, *et al.* (2000) Definition of the *attI1* site of class 1 integrons. *Microbiology* 146(11):2855-2864.
18. Lévesque C, Brassard S, Lapointe J, & Roy PH (1994) Diversity and relative strength of tandem promoters for the antibiotic-resistance genes of several integron. *Gene* 142(1):49-54.
19. Hall RM, Brookes DE, & Stokes H (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol* 5(8):1941-1959.
20. Stokes H, O'gorman D, Recchia GD, Parsekhian M, & Hall RM (1997) Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* 26(4):731-745.
21. Gillings MR (2014) Integrons: past, present, and future. *Microbiol Mol Biol Rev* 78(2):257-277.
22. Bouvier M, Ducos-Galand M, Loot C, Bikard D, & Mazel D (2009) Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genetics* 5(9):e1000632.
23. Loot C, Bikard D, Rachlin A, & Mazel D (2010) Cellular pathways controlling integron cassette site folding. *The EMBO Journal* 29(15):2623-2634.
24. Macdonald D, Demarre G, Bouvier M, Mazel D, & Gopaul DN (2006) Structural basis for broad DNA-specificity in integron recombination. *Nature* 440(7088):1157-1162.

25. Nivina A, Escudero JA, Vit C, Mazel D, & Loot C (2016) 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(16):7792-7803.
26. Bouvier M, Demarre G, & Mazel D (2005) Integron cassette insertion: a recombination process involving a folded single strand substrate. *The EMBO Journal* 24(24):4356-4367.
27. Johansson C, Kamali-Moghaddam M, & Sundström L (2004) Integron integrase binds to bulged hairpin DNA. *Nucleic Acids Res* 32(13):4033-4043.
28. Loot C, Ducos-Galand M, Escudero JA, Bouvier M, & Mazel D (2012) Replicative resolution of integron cassette insertion. *Nucleic Acids Res* 40(17):8361-8370.
29. Meinke G, Bohm A, Hauber J, Pisabarro MT, & Buchholz F (2016) Cre recombinase and other tyrosine recombinases. *Chem Rev* 116(20):12785-12820.
30. Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18(6):292-298.
31. Magadum S, Banerjee U, Murugan P, Gangapur D, & Ravikesavan R (2013) Gene duplication as a major force in evolution. *Journal of Genetics* 92(1):155-161.
32. Gestal AM, Stokes H, Partridge SR, & Hall RM (2005) Recombination between the *dfrA12-orfF-aadA2* cassette array and an *aadA1* gene cassette creates a hybrid cassette, *aadA8b*. *Antimicrob Agents Chemother* 49(11):4771-4774.
33. Recchia GD & Hall RM (1997) Origins of the mobile gene cassettes found in integrons. *Trends Microbiol* 5(10):389-394.
34. Rowe-Magnus DA & Mazel D (2002) The role of integrons in antibiotic resistance gene capture. *Int J Med Microbiol* 292(2):115-125.
35. Fluit A & Schmitz FJ (2004) Resistance integrons and super-integrons. *Clin Microbiol Infect* 10(4):272-288.
36. Ghaly TM & Gillings MR (2018) Mobile DNAs as ecologically and evolutionarily independent units of life. *Trends Microbiol* In press.
37. Page R & Peti W (2016) Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nat Chem Biol* 12(4):208-214.
38. Novikova O & Belfort M (2017) Mobile group II introns as ancestral eukaryotic elements. *Trends Genet* 33(11):773-783.
39. Dai L & Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30(5):1091-1102.
40. Toor N, Robart AR, Christianson J, & Zimmerly S (2006) Self-splicing of a group IIC intron: 5′ exon recognition and alternative 5′ splicing events implicate the stem–loop motif of a transcriptional terminator. *Nucleic Acids Res* 34(22):6461-6471.
41. Quiroga C, Roy PH, & Centron D (2008) The S. ma. I2 class C group II intron inserts at integron *attC* sites. *Microbiology* 154(5):1341-1353.
42. Centrón D & Roy PH (2002) Presence of a group II intron in a multiresistant *Serratia marcescens* strain that harbors three integrons and a novel gene fusion. *Antimicrob Agents Chemother* 46(5):1402-1409.
43. Léon G & Roy PH (2003) Excision and integration of cassettes by an integron integrase of *Nitrosomonas europaea*. *J Bacteriol* 185(6):2036-2041.
44. Sunde M (2005) Class I integron with a group II intron detected in an *Escherichia coli* strain from a free-range reindeer. *Antimicrob Agents Chemother* 49(6):2512-2514.
45. Koenig JE, *et al.* (2008) Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol* 10(4):1024-1038.
46. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, & Mazel D (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the *Vibrionaceae*. *Genome Res* 13(3):428-442.
47. Smith AB & Siebeling RJ (2003) Identification of genetic loci required for capsular expression in *Vibrio vulnificus*. *Infect Immun* 71(3):1091-1097.

48. Robinson A, *et al.* (2008) Structural genomics of the bacterial mobile metagenome: an overview. *Structural Proteomics: High-Throughput Methods*, eds Kobe B, Guss M, & Huber T (Humana Press, Totowa, NJ), pp 589-595.
49. Chang F-Y, Ternei MA, Calle PY, & Brady SF (2015) Targeted metagenomics: finding rare tryptophan dimer natural products in the environment. *J Am Chem Soc* 137(18):6044-6052.
50. Harrington E, *et al.* (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* 104(35):13913-13918.
51. Stokes HW, *et al.* (2001) Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol* 67(11):5240-5246.
52. Koenig JE, *et al.* (2009) Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney Tar Ponds. *PLoS One* 4(4):e5276.
53. Michael CA, *et al.* (2004) Mobile gene cassettes: a fundamental resource for bacterial evolution. *The American Naturalist* 164(1):1-12.
54. Gillings MR, Holley MP, & Stokes H (2009) Evidence for dynamic exchange of *qac* gene cassettes between class 1 integrons and other integrons in freshwater biofilms. *FEMS Microbiol Lett* 296(2):282-288.
55. Elsaied H, *et al.* (2007) Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environ Microbiol* 9(9):2298-2312.
56. Gillings MR, Holley MP, Stokes H, & Holmes AJ (2005) Integrons in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* 102(12):4419-4424.
57. Elsaied H, *et al.* (2011) Marine integrons containing novel integrase genes, attachment sites, *attI*, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *The ISME Journal* 5(7):1162-1177.
58. Goldstein C, *et al.* (2001) Incidence of class 1 and 2 integrases in clinical and commensal bacteria from livestock, companion animals, and exotics. *Antimicrob Agents Chemother* 45(3):723-726.
59. Bailey JK, Pinyon JL, Anantham S, & Hall RM (2010) Commensal *Escherichia coli* of healthy humans: a reservoir for antibiotic-resistance determinants. *J Med Microbiol* 59(11):1331-1339.

Chapter 2 –The biogeography of integron gene cassettes: A window into the protein universe

This chapter is written as a manuscript for submission to *Proceedings of the National Academy of Sciences* and is formatted accordingly. Figures have been inserted within the text at appropriate positions to allow ease of reading and comprehension. Line numbers within the manuscript have been removed to conform with the rest of the thesis.

The biogeography of integron gene cassettes: A window into the protein universe

Timothy M. Ghaly[a,1], Jemma L. Geoghegan[a], John Alroy[a], Michael R. Gillings[a]

[a]Department of Biological Science, Macquarie University, NSW 2109, Australia.

[1]Corresponding author: Timothy Ghaly, tel. +61 (02) 9850 6977, email timothy.ghaly@mq.edu.au.

Short title: Cassettes encode diverse and novel proteins

Classification: Biological Sciences: Microbiology, Ecology and Genetics

Keywords: gene cassette metagenome; protein discovery; microbial diversity; bacterial adaptation; mobile DNA

Abstract

One of the goals of genome sequencing is to assemble a catalogue of all existing protein folds. A global representation of all proteins is likely to be achieved long before we have sequenced Earth's total biodiversity. However, current sequencing methods are inefficient ways to capture the vast pool of novel proteins encoded by rare genes. Integron gene cassettes may provide an alternate and rapid means of assessing the diversity of the protein universe. Here, we sequenced whole gene cassettes from soils sampled across Australia and Antarctica. We recovered 44,970 cassettes that encoded 27,215 unique proteins, representing an order of magnitude more cassettes than previous sequencing efforts. We found that cassettes have an extremely high local richness, with estimates ranging from 4,000 to 18,000 unique cassettes per 0.3 grams of soil. We show that gene cassettes exhibit a rapid spatial turnover, have a heterogeneous distribution across space, and 84% encoded unknown proteins, 64% of which are undocumented in existing databases. Our findings provide useful insights into the vast array of traits available to bacteria with access to the gene cassette metagenome. We also highlight the cassette metagenome as a huge untapped resource that sheds light on the dark matter of the protein universe.

Significance statement

Novel protein discovery is exhibiting diminishing returns from genomic and metagenomic analyses. Consequently, the size and diversity of the protein universe remains unknown. Here, we show that by targeting environmental gene cassettes from integrons, we can raise the discovery rate of entirely novel protein-encoding genes to more than 50% of recovered sequences. We show that this resource is extremely rich and that cassettes have a rapid spatial turnover. This diversity, spatial heterogeneity, and novelty highlight the gene cassette metagenome as a vast resource for bacterial evolution. Their recovery efficiently illuminates the dark matter of the protein universe. Targeted sequencing of integron gene cassettes, as proposed here, could facilitate the complete mapping of the global protein-fold repertoire.

Main Text

Our understanding of the protein universe relies on existing data from sequencing projects. A comprehensive protein catalogue could be assembled long before we have sequenced the global biome (1). This is primarily due to the inherent similarity between proteins, resulting from a shared evolutionary origin (1). Consequently, a global representation of the entire protein-fold repertoire, which must ultimately be finite (2), may be within reach.

Current sequencing methods, however, provide diminishing returns in terms of protein discovery (1, 3). Specifically, whole genome and environmental metagenomic data often fail to capture the vast pool of novel proteins that are encoded by rare genes. Such rare gene clusters are unlikely to be represented within environmental DNA libraries (4). To overcome this limitation, targeting rare genes will enhance protein discovery. Integron gene cassettes can be useful for this purpose since they are widespread in nature, and largely encode novel proteins (5).

Integrons are bacterial genetic elements that can capture, rearrange, and express mobile gene cassettes. They are ancient elements that act as hotspots for generating genomic complexity and adaptive phenotypes. Integrons and their gene cassettes have recently played a significant role in disseminating antibiotic resistance among pathogens (6, 7). However, little is known about gene cassettes outside of clinical settings, even though we know they play an important role in bacterial adaptation and genome evolution (8, 9). All integrons of clinical relevance (classes 1 to 5) are embedded within mobile DNA elements, which provides a point of distinction from chromosomal integrons that lack this mobility. In general, mobile integrons carry few cassettes, often encoding resistance to antibiotics. Chromosomal integrons, which include hundreds of integron classes, generally carry a much larger number of gene cassettes of mostly unknown function. Some *Vibrio* species carry hundreds of cassettes within a single integron (8, 10).

Gene cassettes are ubiquitous, having been recovered from every environment surveyed, including soil and aquatic sediments, the microbiota of plants and animals, aquatic biofilms, seawater, and deep-sea hydrothermal vents (11-15). In these locations, the gene cassette metagenome acts as a vast, floating resource for bacterial evolution. However, the extent of genetic diversity and gene richness captured by the gene cassette metagenome remains unknown. In general, 65% to 80% of open reading frames (ORFs) in cassettes have no known homologs (16-18). Exploring the nature and ecology of the cassette metagenome is a priority for better understanding bacterial diversity and integron-mediated adaptation.

Gene cassettes are flanked by conserved recombination sites (*attC*) that can be targeted by PCR, allowing entire cassettes to be recovered from environmental DNA (5, 17-19). This approach facilitates sequence-independent recovery of diverse ORFs present in bacterial communities. This has a two-fold benefit: first, the diversity, richness, and biogeography of gene cassettes can be

explored; and second, novel proteins may be discovered at an unprecedented rate, given that cassettes commonly encode proteins with unknown functions. The gene cassette metagenome is therefore a key tool for exploring the protein universe, and for asking critical questions about the biogeography of functional gene diversity.

Here, we examine the spatial distribution of mobile gene cassettes across diverse sites in Australia and Antarctica (Fig. 1; Table 1). We find that gene cassettes exhibit extremely high local richness, largely encode novel proteins, have a rapid spatial turnover, and show great spatial heterogeneity. We show that the gene cassette metagenome carries a vast pool of genetic diversity and highlight its potential use in sampling the protein universe.

Fig. 1. Sampling locations. Soil samples were collected from sites in Australia and Antarctica. 24 samples were collected from Sturt National Park (purple square), 11 from Macquarie University campus (blue circle), and 13 from Herring Island (red triangle).

Table 1. Location and description of sampling sites.

Site	*Site description*	*GPS coordinates*	*Number of samples*
Macquarie University campus (NSW)	Urban parkland	33°46'18.26"S 151°6'39.89"E	11
Herring Island (Antarctica)	Antarctic soil	66°24'36.76"S 110°39'17.40"E	13
Rodges (Sturt National Park, NSW)	Semi-arid sandy soil	29°7'1.67"S 141°36'32.25"E	6
Corner (Sturt National Park, NSW)	Semi-arid sandy soil	29°1'29.11"S 141°10'40.28"E	6
Olive Downs (Sturt National Park, NSW)	Semi-arid rocky soil	29°6'36.38"S 141°57'7.99"E	6
Pulgamurtie (Sturt National Park, NSW)	Semi-arid rocky soil	29°5'4.47"S 141°38'9.80"E	6

Results and Discussion

Gene cassette abundance and richness

We sequenced 44,970 gene cassettes from the 48 soil samples listed in Table 1. This represents the largest gene cassette library currently available. The cassette library encoded 27,215 unique proteins (see Methods section for redundancy criteria).

Gene cassettes had extremely high local richness. Chao 1 (20) and squares (21) estimates (using 100 randomizations of the data) predict 4,000 to 18,000 unique cassettes in any one 0.3 g soil sample (Fig. S1-S3).

Any of these gene cassettes could potentially be inserted into any of the hundreds of classes of integron that have now been described (19, 22, 23). This highlights the role of the gene cassette metagenome in facilitating genomic complexity and adaptation in bacterial communities. These numbers are likely to be underestimates of cassette diversity due to primer bias during PCR amplification. The primer pair used here was originally designed against a database that largely contained *attC* sites from antibiotic resistance gene cassettes found on class 1 integrons (5). It is clear that the primer set does not cover the full diversity of this family of recombination sites. Nevertheless, it provides the best richness estimate of gene cassettes to date.

Spatial distribution of gene cassettes

Our data show that gene cassettes exhibit significant heterogeneity across space (Fig. 2). Individual gene cassettes are predominantly rare, with the majority of cassettes occurring at low abundance and with very localized occurrence (Fig. 2). Very few cassettes occurred at high abundance. Even

high abundance cassettes predominantly exhibited a restricted distribution across the sample space (Fig. 2). This supports the idea that the generation of gene cassettes is a dynamic and universal process, occurring continuously at any given location in a landscape. Under this hypothesis, the majority of cassettes are likely to be short-lived, occur at low abundance, and possess very limited spatial distributions. However, in the rare case where a gene cassette provides a substantial selective advantage, it can rapidly increase in both abundance and distribution.

The majority of cassette types occurred in only one sample (Fig. 3). Examination of cassettes present in more than one sample revealed that as occurrence in multiple samples increased, the numbers of individual cassettes with that prevalence rapidly decreased (Fig. 3). Interestingly, however, the number of cassettes present in 80% – 100% of samples increased again (Fig. 3). This pattern was found when examining each environment type separately (Fig. 3a–c), or when pooled (Fig. 3d).

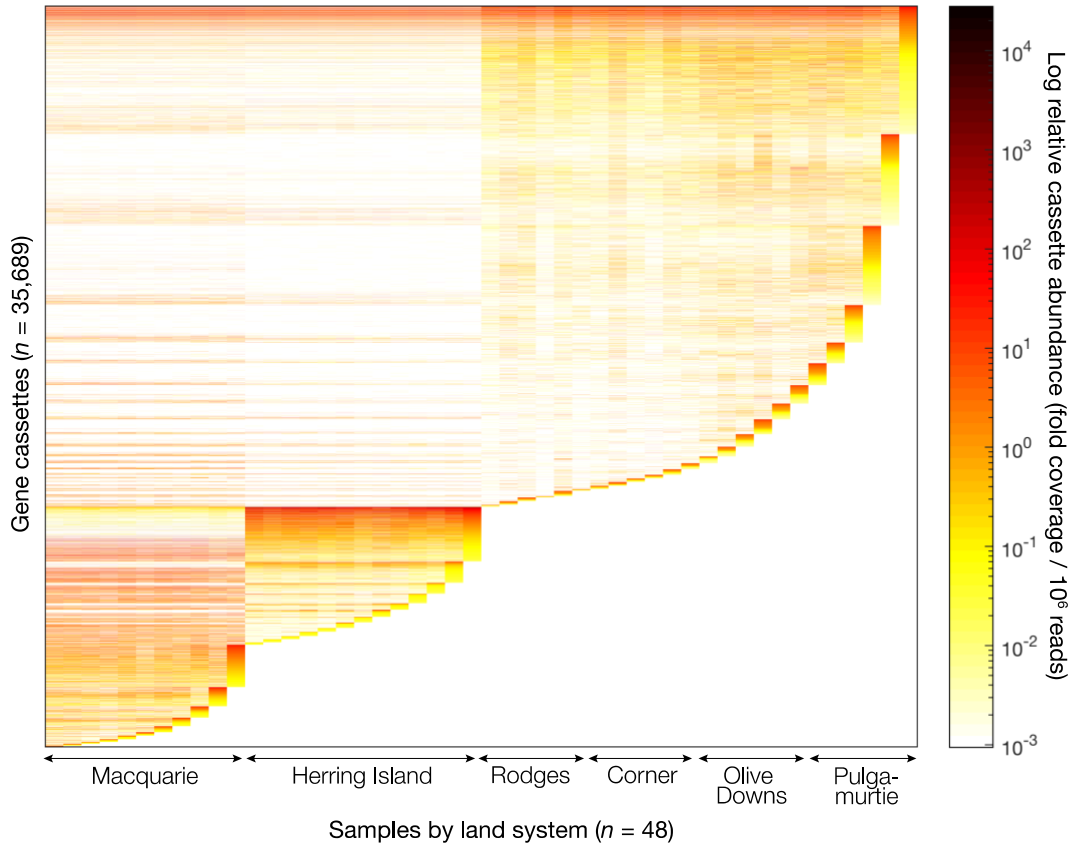

Fig. 2. Relative abundance of gene cassettes. Abundance of each cassette was defined as fold-coverage per million Illumina sequence reads (colour scale bar). The 48 samples include 11 from Macquarie University, Australia; 13 from Herring Island, Antarctica; and 24 from Sturt National Park, Australia (comprises Rodges, Corner, Olive Downs and Pulgamurtie). Note that many gene cassettes reach high local abundance, but are rare or absent in all other samples.

A small subset of cassettes was found in all samples across all sites (Fig. 3d). Of these, 66% encoded entirely novel proteins, while 21% did have homologs, but with no identified function. The remaining cassettes encoded a functionally diverse set of proteins. These included proteins involved in virulence, cell membrane association, protein secretion, signal transduction, defence mechanisms, transport, energy production, protein modification, DNA recombination, and DNA repair and ligation, and even included a CRISPR-associated helicase. For a full list of functions associated with these ubiquitous cassettes, see Table S1.

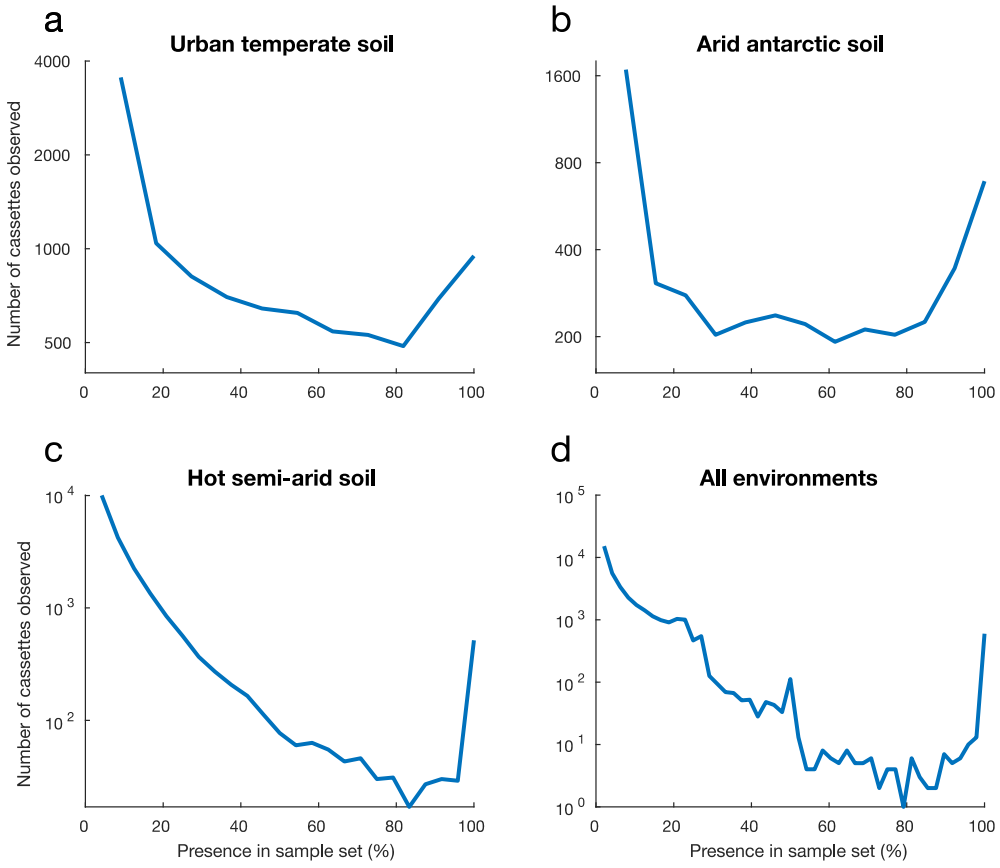

Fig. 3. Spatial distribution of gene cassettes. Number of samples where individual gene cassettes were observed expressed as a percentage of the total number of samples. Samples were divided into different environment types (a-c), or pooled for all environment types (d). The majority of cassettes are found in a few, closely adjacent sampling points, while a subset of cassettes are found in all locations.

We suggest that this distribution pattern can be explained by a combination of three factors: selection, dispersal, and time. We propose that a gene cassette is locally generated within a single cell. This is likely to be the present state of most gene cassettes on the planet, existing as a single copy, or very few copies, at a single location on the globe. If, however, a newly generated cassette provides a significant selective advantage, its abundance will increase, initially through clonal

expansion. Its presence may then become detectable across multiple samples within the same locality as selection increases a cassette's abundance and prevalence.

Over larger spatial scales, dispersal ability must play a major role for the distribution of gene cassettes. Integron gene cassettes are mobile genes, and thus have the potential to disperse through repeated horizontal transfer events. This is most apparent for cassettes found on mobile integrons, which are embedded within transposons, plasmids and integrative-conjugative elements (ICEs) (9, 11). The most notable examples include antibiotic resistance cassettes, which have successfully disseminated into all regions of the globe via horizontal gene transfer of whole integrons, and of the gene cassettes themselves (7, 24). However, most classes of integrons are located on bacterial chromosomes, where horizontal transfer is a much rarer occurrence. Consequently, they depend on vertical transfer within shorter timeframes (9, 25). Gene cassettes can also disperse along with their bacterial hosts. Cassettes present in bacteria with high dispersal abilities, such as planktonic, or spore-forming bacteria, or in bacteria translocated by anthropogenic activities, such as via waste disposal, tourism, or global transport (26), by proxy, will also have an increased capacity for dispersal.

Colonisation of a wide range of environments is dependent on both the ability to disperse and the ability to confer a selective advantage in new locations. This limits the number of individual gene cassette types that can occur at multiple locations. Conversely, there are a number of cassettes that can be found in all samples, regardless of location, suggesting that these cassettes confer phenotypes with general utility (Fig. 3). Gene cassettes that are able to provide a selective advantage in diverse types of environments are also likely to provide an advantage across long temporal scales. Therefore, the number of cassettes that occur ubiquitously will likely accumulate over time.

Gene cassettes that can disperse across global scales and provide a universal selective advantage are likely to be present everywhere. Indeed, some gene cassettes that occurred in every sample in the present study have also been captured using the same PCR-system in additional sites from Australia, Canada, and France (18, 19, 27, 28).

Spatial turnover of gene cassettes

The similarity of gene cassette composition decayed significantly with distance ($r^2 = 0.733$, $P < 0.0001$). Pairwise similarity between samples dropped to 0.1% – 10% beyond an inter-sample distance of 100 metres. This indicates an extraordinarily rapid rate of spatial turnover of gene cassette composition. Regression of probit-transformed similarity against log-transformed distance through all of the data yields a slope of -0.343 (Fig. 4), with a 95% CI between -0.360 and -0.326. The rate of decay over distance varied between regions (Fig. 4), with individual slopes ranging from

-0.05 to -0.15. This suggests that site ecology affects the degree of spatial turnover, and that rates of cassette dispersal are dependent on the spatial scale being examined. Given the high local richness and rapid turnover through space, the size of the gene cassette metagenome is likely to be extensive. It thus represents a vast floating genome that can be drawn upon by diverse bacteria, and that could be exploited as a new source of gene diversity.

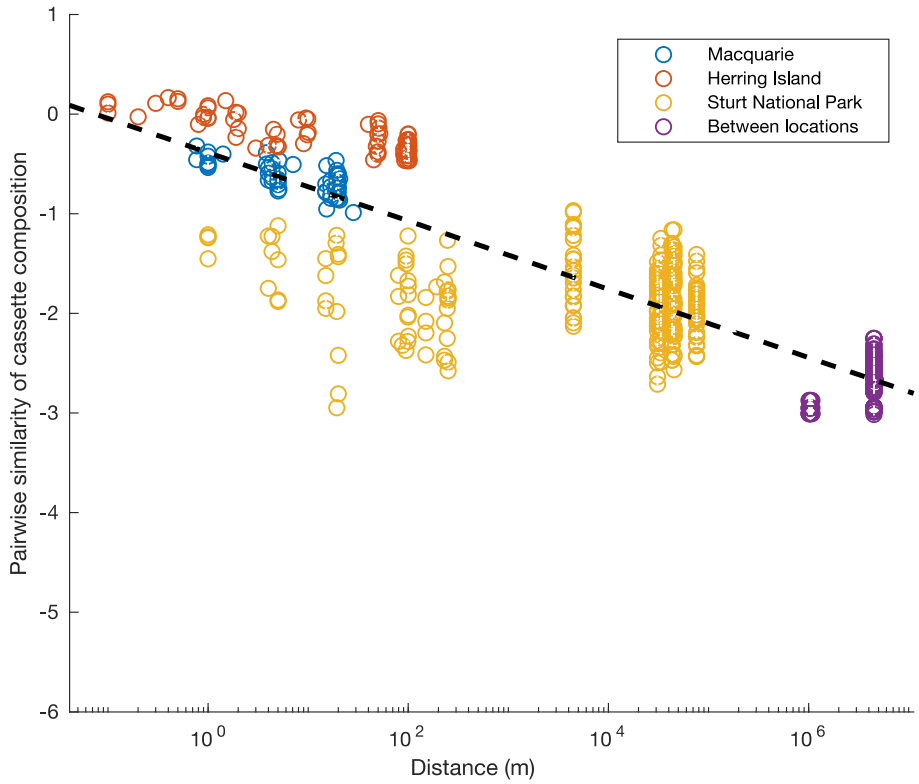

Fig. 4. Probit-transformed pairwise similarity of gene cassette content over log-transformed distances. Across all data, pairwise similarities for gene cassette composition (Forbes index) decline significantly for sampling points separated by 100 metres or more. Rates of distance decay vary between environment types.

Functional novelty of gene cassettes

In addition, our results suggest that the gene cassette metagenome may provide a significant untapped resource for protein discovery. This is highlighted by the extremely high diversity at a given sampling point, as well as the functional novelty of uncovered cassettes. Approximately 54% of the cassettes we recovered encoded proteins with no known homologs, and a further 30% had homologs, but were of unknown function. This is comparable to a previous analysis of gene cassettes from *Vibrio*, which found 78% were either novel or of unknown function (16), and other metagenomic approaches, which determined 78% – 80% of cassettes encoded novel proteins (17, 18). Targeting gene cassettes for sequencing could therefore provide an unprecedented capacity for protein discovery.

The proportion of novel proteins with no known homologs recovered by the methodology presented here is much greater than existing methods. Specific functions can be assigned to 76% of ORFs predicted from environmental metagenomic data, and 83% for completely sequenced genomes (3). These proportions increase to 83% and 86%, respectively, when non-specific functions are considered (3). Our PCR-based approach thus represents an efficient strategy to prospect for novel ORFs in the unknown regions of the protein universe.

Methods

Isolation of environmental DNA

Soil samples were collected from a 1-cm-depth layer from two locations in Australia (Sturt National Park, NSW; and Macquarie University, NSW) and one Antarctic location (Herring Island) (Fig. 1; Table 1). Details of sample collection have been previously described for Sturt National Park (29, 30), which comprised of two semi-arid rocky land systems (Pulgamurite and Olive Downs) and two semi-arid sandy land systems (Rodges and Corner). Sampling at Macquarie University involved three transects within a 20-m grid at 0°, 45°, and 90° angles. Soil samples were collected at 0 m, 1 m, 5 m, and 20 m along the transects. The Herring Island samples were collected on a linear transect at 0 m, 0.1 m, 0.2 m, 0.5 m, 1 m, 2 m, 5 m, 10 m, 50 m, 100 m, 100.1 m, 101 m, and 102 m. DNA was isolated from 0.3 g of each soil sample by previously described bead-beating methods (31).

PCR amplification and DNA sequencing

To amplify complete integron gene cassettes, the primers HS286 and HS287 were used (5). These primers target the conserved sequences for integron *attC* recombination sites, which flank integron-associated gene cassettes. 1 μL of DNA was PCR amplified using GoTaq white (Promega, Madison, WI, USA) and Genereleaser™ (Bioventures, Murfreesboro, TN, USA) as per the manufacturer's protocol. The following thermal cycling program was used: 94°C for 3 min for 1 cycle; 94°C for 30 s, 55°C for 1 min, 72°C for 2 min 30 s, 72°C for 5 min for 35 cycles; and a final cycle at 72°C for 5 min. PCR efficiency was assessed using 1% agarose gel electrophoresis.

In total, 48 samples were selected for sequencing based on the strength of the banding patterns of the electrophoresed PCR products. Twenty-four samples were selected from Sturt National Park, 11 samples from Macquarie University, and 13 samples from Herring Island. Prior to sequencing, PCR products were purified using ExoSAP-IT (Affymetrix, Santa Clara, CA, USA) as per the manufacturer's instructions. All PCR products were sequenced on an Illumina HiSeq2500 platform in two runs (24 samples multiplexed) at the Macrogen sequencing facility (Seoul, South Korea).

DNA sequence assembly and gene cassette annotation

Illumina paired-end (101 bp) reads were first trimmed using Sickle (v1.33) (32) to reach a minimum quality score of 21. Reads shorter than 20 bp after trimming were discarded. If one end of the paired reads had acceptable quality, it was used as a single read. The quality of reads was then inspected using FastQC (v0.11.4) (33). Paired-end reads and single reads were assembled using the Megahit Metagenomic Assembler (v1.1.1) (34).

To ensure that the assembled contigs represented real integron gene cassettes, sequences that were not flanked by both PCR primers were removed using the BBMap package (v35.x) (35). Contigs that were flanked by at least 9 bp from the 3'end of both forward and reverse primers were accepted. After filtering, 60 contigs were randomly sampled for manual inspection, of which, 100% could be identified as putative integron gene cassettes. This was assessed based on the length of the contig (~200 – 1000 bp), and possession of the conserved *attC* regions at each terminus (5).

Genes and translated proteins were called from the filtered contigs using Prokka (v1.12-beta) (36). To measure the total number of unique protein-coding gene cassettes, all translated proteins were compiled into a non-redundant database using CD-HIT (v4.6) (37, 38). A sequence identity cut-off of 97% and minimal alignment coverage greater than 97% for the longer sequence was implemented during clustering.

Pairwise protein alignments

A total of 1,128 pairwise protein alignments between the 48 metagenomic samples were made using DIAMOND (v0.8.33.95) (39). The similarity between any two samples was measured using a rescaled Forbes index (40). The Forbes index was selected as it is less downward-biased relative to other pairwise similarity coefficients for highly diverse datasets (40). Within a pairwise comparison, any two amino acid sequences were considered the same protein according to the above-mentioned redundancy criteria.

References

1. Chubb D, Jefferys BR, Sternberg MJ, & Kelley LA (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics* 26(21):2664-2671.
2. Koonin EV, Wolf YI, & Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420(6912):218-223.
3. Harrington E, *et al.* (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* 104(35):13913-13918.
4. Chang F-Y, Ternei MA, Calle PY, & Brady SF (2015) Targeted metagenomics: finding rare tryptophan dimer natural products in the environment. *J Am Chem Soc* 137(18):6044-6052.
5. Stokes HW, *et al.* (2001) Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. *Appl Environ Microbiol* 67(11):5240-5246.
6. Gillings M, *et al.* (2008) The evolution of class 1 integrons and the rise of antibiotic resistance. *J Bacteriol* 190(14):5095-5100.
7. Partridge SR, Tsafnat G, Coiera E, & Iredell JR (2009) Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 33(4):757-784.
8. Escudero JA, Loot C, Nivina A, & Mazel D (2015) The integron: Adaptation on demand. *Microbiology Spectrum* 3(2):MDNA3-0019-2014.
9. Mazel D (2006) Integrons: agents of bacterial evolution. *Nat Rev Microbiol* 4(8):608-620.
10. Chen C-Y, *et al.* (2003) Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res* 13(12):2577-2587.
11. Ghaly TM, Chow L, Asher AJ, Waldron LS, & Gillings MR (2017) Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS One* 12(6):e0179169.
12. Gillings MR, Holley MP, & Stokes H (2009) Evidence for dynamic exchange of *qac* gene cassettes between class 1 integrons and other integrons in freshwater biofilms. *FEMS Microbiol Lett* 296(2):282-288.
13. Elsaied H, *et al.* (2007) Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. *Environ Microbiol* 9(9):2298-2312.
14. Gillings MR, Holley MP, Stokes H, & Holmes AJ (2005) Integrons in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* 102(12):4419-4424.
15. Elsaied H, *et al.* (2011) Marine integrons containing novel integrase genes, attachment sites, *attI*, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *The ISME Journal* 5(7):1162-1177.
16. Boucher Y, Labbate M, Koenig JE, & Stokes H (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol* 15(7):301-309.
17. Koenig JE, *et al.* (2008) Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol* 10(4):1024-1038.
18. Koenig JE, *et al.* (2009) Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney tar ponds. *PLoS One* 4(4):e5276.
19. Abella J, Fahy A, Duran R, & Cagnon C (2015) Integron diversity in bacterial communities of freshwater sediments at different contamination levels. *FEMS Microbiol Ecol* 91(12):fiv140-fiv140.
20. Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11(4):265-270.
21. Alroy J (2018) Limits to species richness in terrestrial communities. *Ecol Lett* In press.
22. Cambray G, Guerout A-M, & Mazel D (2010) Integrons. *Annu Rev Genet* 44(1):141-166.
23. Boucher Y, *et al.* (2011) Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. *MBio* 2(2):e00335-00310.
24. Gillings MR (2017) Class 1 integrons as invasive species. *Curr Opin Microbiol* 38:10-15.
25. Gillings MR (2014) Integrons: past, present, and future. *Microbiol Mol Biol Rev* 78(2):257-277.
26. Zhu Y-G, *et al.* (2017) Microbial mass movements. *Science* 357(6356):1099-1100.

27. Michael CA, *et al.* (2004) Mobile gene cassettes: a fundamental resource for bacterial evolution. *The American Naturalist* 164(1):1-12.
28. Holmes AJ, *et al.* (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol* 5(5):383-394.
29. Oliver I, *et al.* (2004) Land systems as surrogates for biodiversity in conservation planning. *Ecol Appl* 14(2):485-503.
30. Green JL, *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature* 432(7018):747-750.
31. Yeates C, Gillings M, Davison A, Altavilla N, & Veal D (1998) Methods for microbial DNA extraction from soil for PCR amplification. *Biol Proced Online* 1(1):40-47.
32. Joshi N & Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].
33. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data (Version 0.11.4) [Software].
34. Li D, Liu C-M, Luo R, Sadakane K, & Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10):1674-1676.
35. Bushnell B (2014) BBMap: a fast, accurate, splice-aware aligner. (Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US)).
36. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068-2069.
37. Li W, Jaroszewski L, & Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17(3):282-283.
38. Li W, Jaroszewski L, & Godzik A (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18(1):77-82.
39. Buchfink B, Xie C, & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12(1):59-60.
40. Alroy J (2015) A new twist on a very old binary similarity coefficient. *Ecology* 96(2):575-586.

Chapter 3 – Final discussion and concluding remarks

In this thesis, I have shown that integron gene cassettes have an extremely high local richness, ranging from 4,000 to 18,000 cassettes per 0.3 grams of soil, have a rapid turnover through space, exhibit significant spatial heterogeneity, and largely encode novel proteins. Several important implications arise from these findings. First, this study provides the most in-depth analysis to date of the ecology of the gene cassette metagenome; second, the extent to which gene cassettes can facilitate genome complexity and drive bacterial evolution is shown; and finally, these results highlight the potential use of integron gene cassettes in accelerating the discovery of novel proteins.

Gene cassette ecology

Gene cassettes are predominantly rare. This was evident from the majority of cassettes being present in only one soil sample. Examination of cassettes present in more than one sample revealed that as occurrence in multiple samples increased, the numbers of individual cassettes with that prevalence rapidly decreased. This pattern fits the hypothesis that gene cassettes are ubiquitous and are generated at any given locality across a landscape. However, somewhat more surprisingly, the number of gene cassettes that were present in 80% to 100% of samples increased again. This indicates that a subset of the gene cassette pool is extremely common, or even ubiquitous, within the global soil microbiome. While most gene cassettes have a limited spatial distribution, a select few must confer a universal advantage and thus have the capacity to disperse and persist in a broad range of environment types across the globe. The functions of these cassettes are clearly worth further investigation. The widespread distribution of these cassettes may be a result of their mobile nature, facilitating repeated horizontal transmission events, or as a consequence of host cell dispersal over time.

Cassettes that are able to provide a universal advantage are also likely to provide an advantage across long temporal scales. Therefore, the number of cassettes that occur ubiquitously likely accumulate over time. This opens the possibility of a floating genome to be shared among diverse bacteria existing in very different environment types.

It should be noted that among the cassettes detected here, there were no instances of the typical antibiotic resistance cassettes that are commonly associated with clinical integrons (1). This further demonstrates the rarity of any particular cassette type within an environmental sample. It is likely that the cassettes currently driving the global resistance crisis have become common in clinical contexts precisely because of the drivers outlined above. Under antibiotic selection, capture of resistance cassettes conferred a significant advantage, and their location on mobile elements then

allowed rapid dispersal between cells. Colonisation of human commensal and pathogenic bacteria then provided a means for global dissemination over the last 80 years (2, 3).

The extremely high local richness of gene cassettes, in conjunction with their spatial heterogeneity and rapid turnover through space, indicates that the gene cassette metagenome is of immense size. Any bacterium in possession of the gene-capturing integron system has access to the genetic diversity within this metagenome. Since gene cassette acquisition relies more on the conserved structure of *attC* sites, evolutionarily distinct integrons can capture any available gene cassette, which collectively carry very different *attC* sequences (4, 5).

One interesting observation to be made is the disproportionately large number of resistance cassettes that are carried by mobile integrons. Such integrons, embedded within mobile DNA elements, can rapidly spread across broad phylogenetic boundaries with access to a significant portion of the gene cassette metagenome. Indeed, cassettes associated with mobile integrons appear to have been collected from multiple genomic backgrounds. This can be inferred by the inconsistent codon usage of their cassette ORFs, and the vast sequence diversity of their *attC* sites (6). In contrast, chromosomal integrons have cassette arrays with highly similar *attC* sites (7-10). Homologous *attC* sites are species-specific and cluster according to host topology (8, 9). This provides a clear distinction, whereby chromosomal integrons generally recruit gene cassettes of intraspecific origin, while mobile integrons disperse between species sampling cassettes of diverse origins. Despite the heterogenous origins of cassettes associated with mobile integrons, they exhibit remarkably homogenous functionality, with the majority of cassettes conferring resistance to antibiotics.

A likely explanation is that resistance cassettes largely function as single genes, with minimal disturbance to the existing genome. In general, the successful integration of heterologous genes in a new host depends on their interaction with specific components of gene regulatory networks and host physiology (11, 12). Cassettes associated with chromosomal integrons are unlikely to disturb host physiology and metabolism given their intraspecific origin. In mobile integrons, however, heterologous cassettes are much more likely to be disruptive to host fitness. This may explain why the majority of cassettes associated with mobile integrons encode antibiotic-modifying enzymes that function without significant cellular interactions (1). These cassettes largely encode acetyltransferases, β-lactamases and nucleotidyl transferases (1). In contrast, genes conferring antibiotic resistance through regulatory mechanisms are rarely observed within mobile integrons (1). This suggests that mobile integrons would never have proliferated if it were not for the sustained selection pressure from human antibiotic use. Such elements are now being classed as xenogenetic DNAs, which are defined as DNA molecules with a recent origin, and whose assembly and dissemination was driven by human activities (13).

Functional diversity of gene cassettes

Outside of clinical settings, it is clear that the gene cassette metagenome encodes a functionally diverse range of proteins. This is evident from the examination of the 596 gene cassettes that were found in all samples in the present study. General functions could be assigned to less than 13% of these ubiquitous cassette-encoded proteins, yet even these cover a wide range of biological functions. These include virulence factors, cell membrane association, protein secretion, signal transduction, defence mechanisms, transport, energy production, protein modification, DNA recombination, and DNA repair and ligation. Interestingly, one of these ubiquitous cassettes encoded a CRISPR-associated helicase. For a full list of functions associated with these ubiquitous cassettes, see Table S1 (Appendix).

Such functional diversity supports findings from previous studies. Similarly, functions that have been previously assigned to the few chromosomal gene cassettes with characterised homologs also exhibit considerable diversity. These include DNA modification, toxin-antitoxin systems, phage-related functions, isochorismatases, acetyltransferases and virulence factors (6, 14, 15). Further, a number of gene cassettes have had their functions experimentally determined, including restriction or methylation systems, sulfate-binding proteins, lipases, polysaccharide biosynthesis, and dNTP pyrophosphohydrolases (9, 16, 17). To add to this functional repertoire, a large number of cassette-encoded proteins possess conserved domains involved in membrane association and cellular export (8, 15). Such domains signify the role gene cassettes play in mediating interactions between host bacteria and their surrounding environment.

Here, I have shown the extremely high gene richness existing within the cassette metagenome, which, along with its considerable functional diversity, indicate that gene cassettes must play a significant role in driving phenotypic diversity and adaptation in bacteria.

Sampling the protein universe

This broad functional diversity, however, is associated with a small proportion of gene cassette that encode proteins with characterised homologs. This study compiled the largest gene cassette library to date and revealed that 84% of cassettes encoded proteins that were novel or of unknown function. Similarly, previous studies estimated between 78% – 80% of cassettes encoded novel proteins (15, 18). The high number of novel proteins detected, relative to the sequencing effort of this PCR-based approach, presents a promising opportunity for protein discovery.

Our understanding of the protein universe relies on existing data from previous sequencing projects, and a globally representative map should be reached long before we have sequenced Earth's biodiversity (19). Existing sequencing methods, however, provide diminishing returns in

regards to protein discovery (19, 20). To capture the diversity of existing protein families, efforts should focus on the vast pool of novel proteins that are encoded by rare genes. Such rare gene clusters are unlikely to be represented within environmental DNA libraries (21). The overrepresentation of rare and novel proteins encoded by the gene cassette metagenome, as shown in this thesis, indicates that a targeted metagenomic approach via gene cassette PCR can help increase rates of novel sequence accumulation. Furthermore, the high local richness and spatial heterogeneity of gene cassettes implies that repeated cassette sequencing projects would substantially increase the number of known proteins. Consequently, a full representation of the global protein fold repertoire, which must ultimately be finite (22), may be within reach.

In order to enhance our understanding of the gene cassette metagenome, and more generally, the protein universe, further gene cassette sequencing projects would be useful. Additional sampling across greater spatial scales may allow for a global estimate of the size of the gene cassette metagenome. Furthermore, sampling over temporal scales may provide information on the timeframes in which gene cassettes are generated. Such studies are necessary to reveal the ecology and evolution of integron gene cassettes.

References

1. Partridge SR, Tsafnat G, Coiera E, & Iredell JR (2009) Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 33(4):757-784.
2. Gillings M, *et al.* (2008) The evolution of class 1 integrons and the rise of antibiotic resistance. *J Bacteriol* 190(14):5095-5100.
3. Ghaly TM, Chow L, Asher AJ, Waldron LS, & Gillings MR (2017) Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS One* 12(6):e0179169.
4. Bouvier M, Ducos-Galand M, Loot C, Bikard D, & Mazel D (2009) Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genetics* 5(9):e1000632.
5. Loot C, Bikard D, Rachlin A, & Mazel D (2010) Cellular pathways controlling integron cassette site folding. *The EMBO Journal* 29(15):2623-2634.
6. Cambray G, Guerout A-M, & Mazel D (2010) Integrons. *Annu Rev Genet* 44:141-166.
7. Gillings MR, Holley MP, Stokes H, & Holmes AJ (2005) Integrons in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* 102(12):4419-4424.
8. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, & Mazel D (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the *Vibrionaceae*. *Genome Res* 13(3):428-442.
9. Rowe-Magnus DA, *et al.* (2001) The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc Natl Acad Sci U S A* 98(2):652-657.
10. Vaisvila R, Morgan RD, Posfai J, & Raleigh EA (2001) Discovery and distribution of super-integrons among Pseudomonads. *Mol Microbiol* 42(3):587-601.
11. Porse A, Schou TS, Munck C, Ellabaan MMH, & Sommer MOA (2018) Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nature Communications* 9(1):522.
12. Jain R, Rivera MC, & Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 96(7):3801-3806.
13. Gillings MR, Westoby M, & Ghaly TM (2018) Pollutants that replicate: Xenogenetic DNAs. *Trends Microbiol* In press.
14. Boucher Y, Labbate M, Koenig JE, & Stokes H (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol* 15(7):301-309.
15. Koenig JE, *et al.* (2008) Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol* 10(4):1024-1038.
16. Smith AB & Siebeling RJ (2003) Identification of genetic loci required for capsular expression in *Vibrio vulnificus*. *Infect Immun* 71(3):1091-1097.
17. Robinson A, *et al.* (2008) Structural genomics of the bacterial mobile metagenome: an overview. *Structural Proteomics: High-Throughput Methods*, eds Kobe B, Guss M, & Huber T (Humana Press, Totowa, NJ), pp 589-595.
18. Koenig JE, *et al.* (2009) Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney Tar Ponds. *PLoS One* 4(4):e5276.
19. Chubb D, Jefferys BR, Sternberg MJ, & Kelley LA (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. *Bioinformatics* 26(21):2664-2671.
20. Levitt M (2009) Nature of the protein universe. *Proc Natl Acad Sci U S A* 106(27):11079-11084.
21. Chang F-Y, Ternei MA, Calle PY, & Brady SF (2015) Targeted metagenomics: finding rare tryptophan dimer natural products in the environment. *J Am Chem Soc* 137(18):6044-6052.
22. Koonin EV, Wolf YI, & Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420(6912):218-223.

Appendix – Supplementary data

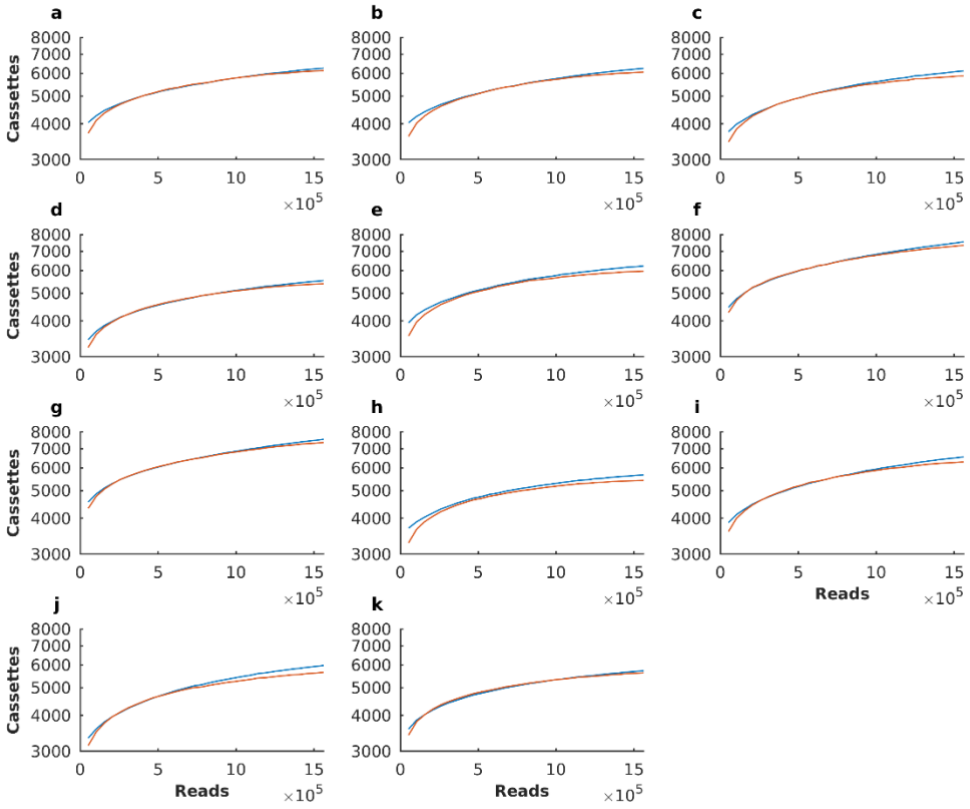

Fig. S1. Richness extrapolation curves for Macquarie University samples. Cassette richness obtained from number of Illumina sequence reads based on Chao 1 (red) and squares (blue) analyses. Plots **a** to **k** correspond to the 11 samples from the Macquarie University sampling site.

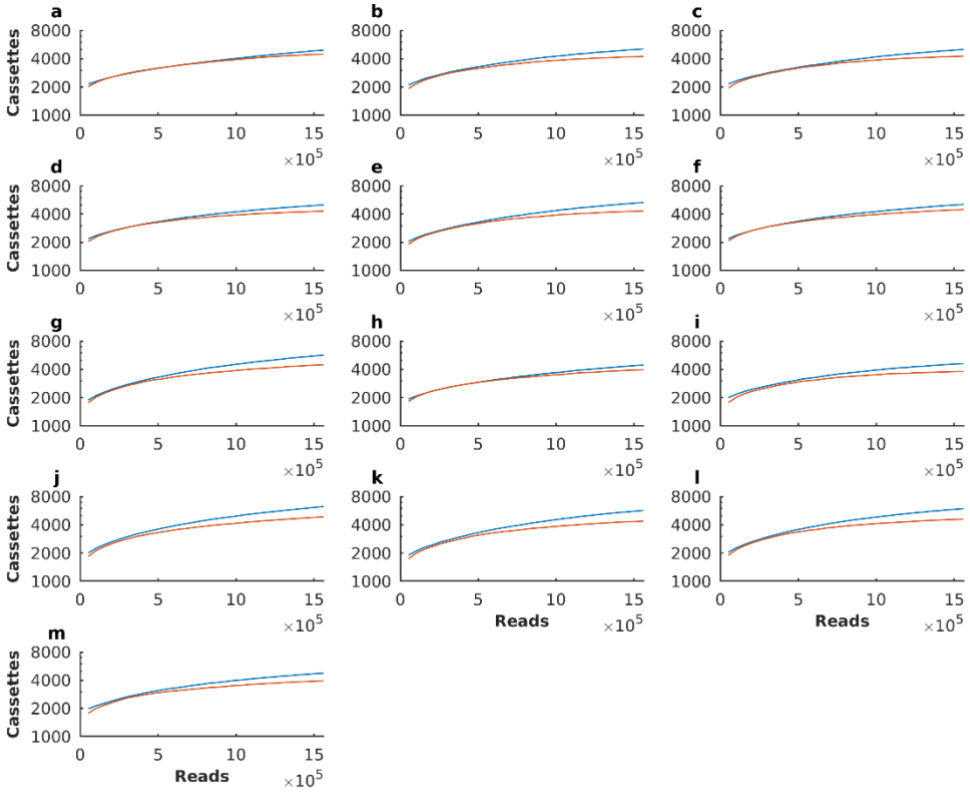

Fig. S2. Richness extrapolation curves for Herring Island samples. Cassette richness obtained from number of Illumina sequence reads based on Chao 1 (red) and squares (blue) analyses. Plots **a** to **m** correspond to the 13 samples from Herring Island, Antarctica.

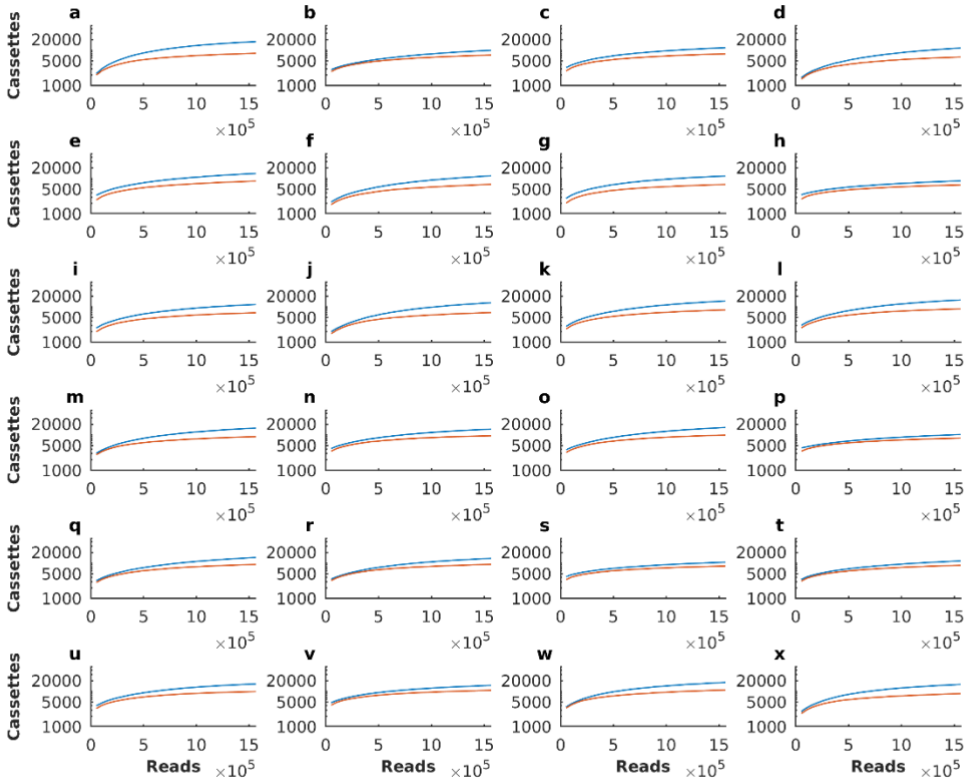

Fig. S3. Richness extrapolation curves for Sturt National Park samples. Cassette richness obtained from number of Illumina sequence reads based on Chao 1 (red) and squares (blue) analyses. Plots **a** to **f** correspond to the 6 samples from Rodges; **g** to **l** correspond to the 6 samples from Corner; **m** to **r** correspond to the 6 samples from Olive Downs; and **s** to **x** correspond to the 6 samples from Pulgamurtie.

Table S1. Proteins encoded by the 596 gene cassettes that were present in all 48 soil samples.

Cassette-encoded protein	*Number of cassettes that encode protein*
Novel hypothetical protein	394
Conserved hypothetical protein	125
Transposase	4
Type II secretion system protein	3
HNH endonuclease	3
Glyoxylase / bleomycin resistance family protein	3
TonB energy transducer	2
Cation transporter	2
GNAT family N-acetyltransferase	2
Barnase inhibitor	2
Cas3, CRISPR-associated helicase	1
Putative membrane protein	1
GFA family protein	1
DNA ligase, LigD	1
2-hydroxy-3-oxopropionate reductase	1
Electron transfer Ferredoxin protein	1
Periplasmic TRAP transporter protein	1
COXI, cytochrome c oxidase subunit I	1
Phosphoribosylglycinamide formyltransferase	1
carbamoyl-phosphate synthase large subunit	1
Glycoside hydrolase superfamily protein	1
Methionine synthase I	1
TP_methylase superfamily protein	1
Dihydrofolate reductase	1
Carboxypeptidase-like, regulatory domain superfamily	1
ACT domain-containing protein	1
Suppressor of fused protein domain protein	1
Virulence factor protein	1
1,4-dihydroxy-2-naphthoyl-CoA synthase	1
Malate dehydrogenase	1
DNA methylase	1
luciferase family oxidoreductase	1
Dehydrogenase	1
glycine/betaine ABC transporter substrate-binding protein	1

Fe-S cluster assembly protein SufB	1
Putative lactoylglutathione lyase	1
NDP-hexose 4-ketoreductase	1
NACHT domain-containing protein	1
Epoxide hydrolase	1
Peptidase	1
Dihydrolipoyl dehydrogenase	1
Homoserine kinase	1
RecF DNA replication and repair protein	1
Sensor histidine kinase	1
Deaminase	1
6-carboxytetrahydropterin synthase	1
Beta-lactamase	1
PKD domain protein	1
ATP-binding protein	1
XRE family transcriptional regulator	1
ATP-dependent Clp protease ATP-binding subunit	1
Putative transcriptional regulator	1
Endolytic transglycosylase	1
Helix-turn-helix transcriptional regulator	1
Modification methylase PaeR7I	1
SnoaL-like domain protein	1
Nucleotidyltransferase	1
DNA primase-like protein	1
N-acetyltransferase	1
MBL fold metallo-hydrolase	1
Carbohydrate ABC transporter permease	1
HAD family hydrolase	1
Glycosyl transferase	1
FAD-binding protein	1
DNA polymerase III alpha subunit	1
VWA domain-containing protein	1