UNIVERSITY OF CALIFORNIA
SANTA CRUZ

COMPUTATIONAL GENOMICS STUDIES OF GENETIC ADAPTATION IN DIFFERENT ENVIRONMENTAL AND ORGANISMAL CONTEXTS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Jay Wook Joong Kim

December 2019

The dissertation of Jay Wook Joong Kim
is approved:

Dr. Manel Camps, Chair

Dr. Richard Ed Green

Dr. Todd M. Lowe

Dr. Karen Ottemann

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Computational genomics studies of genetic adaptation in different environmental and organismal contexts

by

Jay Wook Joong Kim

This dissertation presents a study of global patterns in the distribution of acquired resistance genes in 352 draft genomes of *E. coli* samples from two US West Coast hospitals. A nondeterministic clustering of genomes based on their resistance gene composition identifies two highly successful gene combinations involving β-lactamase and aminoglycoside acetyltransferase genes that largely explain the distribution of ESBL, gentamicin, and tobramycin resistance in these samples. We name these two parallel adaptive solutions "Complementarity Groups 1 and 2" (CG1 and CG2) because we observe functional diversification of genes within these groups, driven by mutual antagonism between genes exhibiting similar resistance profiles. Mutual antagonism extending across groups drives parallel adaptive trajectories. In 761 completely assembled genomes from NCBI, representing a broader range of geographical and ecological sources, we confirm: (1) the prevalence of CG1 and CG2; (2) establish that mutual antagonism is a generalized feature of acquired resistance genes with overlapping function; (3) and verify that the observed gene-to-gene associations correspond to physical linkages. We also find that configurations placing gene pairs in high proximity and on the same strand tend to be more successful. We propose a model that explains these observations, constraining evolution through a combination of physical linkage and mutual antagonism in the context of generalized panmixia. Looking at the genomic context for antibiotic resistance genes in NCBI761, we find mosaic plasmids (with replicons belonging to different incompatibility groups), a complex network of linkages between replicons and resistance (dominated by IncF), and significant gene flow to the chromosome, particularly for ESBLs.

Also, I describe a deep mutational scanning approach for directed evolution of proteins, and the generation of TEM β-lactamase mutant libraries using this approach. Each mutant library captures alternate sequence subspaces in the evolution of extended-spectrum resistance (a gain-of-function), and are generated by leveraging negative epistasis between their respective starting points for directed evolution.

In the future the dataset generated by this approach will enable the study of higher order mutational interactions in the evolution of extended-spectrum resistance.

In addition, two other studies are presented: (1) A comparison of the repeat landscapes in the genomes of 8 ant species highlights the role of transposable element clusters (TE islands) in facilitating the adaptation of an invasive species to new habitats. (2) A comparison of the genomes of three marine Planctomycetes inhabiting the blade of the red alga, *Porphyra umbilicalis*. These three OTUs represent three different genera, and contain large expansions of specific gene families and horizontally acquired genes, which appear to augment their metabolic repertoire for accessing macropolymers in the cell walls of algae, and their mechanisms for stress responses that likely help adaptation to the intertidal zone.

For my immigrant father whose sacrifices have amounted to my American dream.
For my mother, who does not waver and whose support is entirely unconditional.
For my brother, who was my early-life male role model.
For my August Sun, who through no fault of his own has already endured much during his short time with us.

For my mentors, who shared with me their visions, expertise and life lessons.
For my students who continue to inspire me and fill me with purpose.

For my homeless neighbors and other single-serving friends struggling with mental health conditions.

For the 99.9%.
And for Mr. Sanders, who strives to make things better for all of us.

Acknowledgments

The experimental work in Chapter 1 of this thesis was performed in collaboration with Dr. Gerardo Cortés-Cortés, and with the laboratory of Dr. Miriam Barlow at the University of California, Merced. I would like to thank Dr. Gerardo Cortés-Cortés, Anne Courney and Ethan Santos for their work in performing, interpretting and presenting the conjugation and primer-walking assays. I would also like to thank Dr. Portia Mira for her work in culturing and preparing the samples from Dignity Health Mercy Medical Center for sequencing, and Dr. Miriam Barlow for her tireless work in obtaining these samples, and for generously sharing the samples.

Additionally, I would like to acknowledge Dr. Steve Salipante for sharing 39 isolates from the University of Washington, and Dr. Luis Mota-Bravo (University of California, Irvine) for sharing his unpublished conjugation protocols and recipient strains.

Finally, I would like to thank Dr. Manel Camps for advising the work in this thesis, for help formulating ideas, and for help editing the manuscript for publication. I also thank my committee, Drs. Todd Lowe, Ed Green and Karen Ottemann, for their valuable suggestions and general guidance.

This work was funded by CITRIS Seed Funding proposal 2015-324 to Todd Lowe, Manel Camps and Miriam Barlow, by NIAD award 1R41AI122740-01A1 to Manel Camps and Miriam Barlow in partnership with Maverix Genomics, and by a UC MEXUS-CONACYT award 18/19, 19900-443798-A19-0609-002 to Gerardo Cortés-Cortés.

Chapter 1

Mechanisms constraining the acquired multidrug resistance landscape in *E. coli*

1.1 Introduction

The rise of multidrug resistance (MDR) in bacteria poses a serious threat to public health, as multidrug-resistant infection is linked to extended hospitalization and mortality [1–4]. A better understanding of how MDR arises and spreads could be of great assistance in informing strategies to control it.

Our study focuses on extraintestinal pathogenic *E. coli* (ExPEC), which are particularly effective at acquiring new resistance genes [5, 6]. *E. coli* infection is an important threat to human health because of its frequent implication in opportunistic infections, causing urinary tract infections, sepsis and wound infections [7,8].

The evolution of MDR represents a complex process leading to the accumulation of resistance factors in a single bacterial strain [9]. Genetic resistance factors include chromosomal mutations that modify drug target sites, genes encoding drug-inactivating enzymes, and genes encoding multidrug efflux pumps that can export multiple types of antibiotics across bacterial cytoplasmic membranes. Resistance factors can occur in the core genome or in the peripheral genome, which includes extrachromosomal elements and sections of the chromosome that contain a high proportion of mobile elements. Resistance factors in the core genome are more aligned with vertical transmission, while acquired resistance genes in the peripheral genome are primarily encoded in plasmids and can be highly mobile.

The evolution of MDR, especially within the peripheral genome, is facilitated by various mobile genetic elements (MGEs) [10] including transposons and integrons. These elements provide both the means for bacteria to acquire resistance genes from a shared environmental pool [11–13] and to increase genetic diversity by generating different gene combinations (reviewed in [14, 15]). Horizontal gene transfer (HGT) plays a critical role in extending the genetic diversity available by facilitating gene flow across different strains and even across different microorganisms [16,17] and explains how virtually identical plasmids can be found in unrelated bacterial strains [18, 19].

The acquired resistance genes most frequently found in *E. coli* genomes include genes encoding aminoglycoside-modifying enzymes (AMEs), extended-spectrum β-lactamases (ESBLs) and carbapenemases. AMEs can be subdivided into three groups based on their catalytic reaction, namely acetyltransferases (AACs), adenyltransferases (ANTs) or phosphotransferases (APHs) [20]. ESBLs include the CTX-M and CMY groups primarily, and the TEMs, and SHVs [21]. Carbapenemase groups include the NDMs, IMPs, KPCs, VIMs and some OXAs [5].

These genes frequently cluster in large multi-resistance regions (MRRs) [15, 22]. The evolution of MDR is thought to involve a process called genetic capitalism, wherein the presence of acquired antibiotic resistance genes in a given strain facilitates the acquisition of resistance to additional antibiotics [11,23]. According to this model, antibiotic resistance genes increase their frequency in the gene pool vertically (through clonal expansion) and horizontally (largely through conjugational transfer on plasmids). This increased representation in the gene pool improves access to other antibiotic resistance genes for recombination, thus facilitating the generation of larger, more successful gene combinations. The role of recombination is evidenced by the consistent identification of modules of a few genes arranged in the same way and with the same boundaries in different contexts [15, 22, 24].

Genetic capitalism is facilitated by co-selection, a process where one antibiotic indirectly selects for genes conferring resistance to different antibiotics [23]. In MRRs, co-selection is driven by the physical linkage between resistance genes and its efficiency is inversely proportional to the distance between resistance genes. Another mechanism involved in adaptation is mutual antagonism. This mechanism has been extensively studied in model systems involving a single protein such as TEM-1 or phosphoglycerate kinase [25, 26], where it generally involves some incompatibility between key adaptive mutations. In these model systems, mutual antagonisms that arise early during adaptation constrains evolutionary trajectories further down-

stream, a process known as contingency [27].

Here we turned to whole genome sequencing to investigate the contribution of physical linkage and mutual antagonism between resistance genes to the evolution of MDR in the peripheral genome. Previous comparative genomics studies that provide a comprehensive view of the multidrug "resistome" have already reported some frequent gene combinations [7, 28–32], but the overall landscape of these adaptive solutions as well as the role of gene-to-gene interactions in their evolution have not been systematically investigated.

We conducted a case study of 352 clinical isolates of ExPEC from two hospitals on the U.S. West Coast focused on genes contributing to resistance against aminoglycosides and β-lactams, two antibiotic classes frequently used to treat enterobacterial infection [6,33,34]. We identified two highly successful gene combinations that largely explain the distribution of aminoglycoside and β-lactam resistance in the clinical populations we investigated. We also identified antagonistic interactions between genes with similar resistance profiles. These findings were verified against 761 completed *E. coli* genomes from the NCBI database and we used these fully assembled genomes to investigate the proximity between these genes and the fluidity in their linkage.

Based on these observations, we propose a model that explains the evolution and adaptive success of a restricted number of gene combinations. Our model includes a significant role for both physical linkage and mutual antagonism. As part of the model, we propose a process of stochastic fine-tuning by recombination and selection that would lead to the observed trend toward high proximity and co-strandedness between antibiotic resistance genes within adaptive solutions. Looking at the physical linkage between replicons, we confirm extensive replicon mosaicism between IncF replicons and also find mosaicism involving other replicon types. This replicon mosaicism could broaden the compatibility range of the plasmids involved, potentially accelerating the process of MDR adaptation. Individual resistance genes were typically linked to a variety of replicons, although IncF plasmids were particularly central. We also observed significant gene flow to the chromosome, especially for ESBL genes.

1.2 Results and Discussion

1.2.1 Description of whole-genome datasets

Three independent, non-overlapping whole-genome datasets including 1113 *E. coli* genomes were used to support the analyses and conclusions in this study: UW233, DHMMC119, and NCBI761. A summary of the composition of the three datasets is provided in **Table 1.1**.

The UW233 dataset contained 233 ExPEC draft genomes from a published study conducted at the University of Washington [8], representing the unbiased sampling of the clinical strains isolated in a hospital, from patients suffering from sepsis between 2008 and 2013, or from urinary tract infections between 2011 and 2012. This group of samples had a relatively low incidence of ESBL resistance (19.6% of isolates with antibiotic sensitivity data). The DHMMC119 dataset contained 119 draft genomes that we sequenced from a collection of ExPEC isolates from the Dignity Health Mercy Medical Center in Merced, CA, USA. This collection was deliberately enriched for strains exhibiting ESBL resistance (61.6%, of isolates with susceptibility data) and supplemented the low incidence of MDR in UW233 (see Methods).

For statistical analysis, we merged the UW233 and DHMMC119 datasets into USWest352. These genomes were accompanied by clinical data, including results from antibiotic susceptibility testing against a panel of antibiotics, with some differences in panel composition between the two. The antibiotics that we focus on in this study, and their classification are shown in **Table 1.2**. USWest352 strains displayed 32.3% incidence of ESBL resistance, 26.8% incidence of aminoglycoside resistance, and 47.9% incidence of FQN resistance. Thus, USWest352 provided an adequate representation for statistical analysis on the distribution of resistance markers across two hospitals in the U.S. West Coast despite not representing a random sampling of the clinical population.

Table 1.1: Summary of five whole-genome datasets used in this study.

Dataset	Size	Source	Assembly Status	Antibiotic susceptibility	Sampling enrichment	Total MLSTs	% MLST composition (Top 5)
UW233	233	University of Washington	draft	available	none	75	ST131 (15.5%), ST95 (11.6%), ST73 (9.4%) ST127 (6.4%), ST69 (5.2%)
DHMMC119	119	Dignity Health Mercy Medical Center	draft	available	ESBL	22	ST131 (58.0%), ST648 (10.9%), ST44 (5.0%) ST38 (4.2%), ST405 (3.4%)
USWest352	352	UW233, DHMMC119	draft	available	ESBL	84	ST131 (29.8%), ST95 (8.0%), ST73 (6.3%) ST127 (4.8%), ST69 (3.7%)
NCBI761	761	NCBI	complete	not available	none	200	ST10 (10.9%), ST131 (7.4%), ST11 (6.4%) unknown (5.9%), ST167 (2.5%)

The NCBI761 dataset included all 761 completely assembled, non-synthetic *E. coli* genomes available in the NCBI database on April 3, 2019. These genomic sequences provided an independent dataset for verifying and framing our findings in a larger context outside of the two hospitals. In addition, these completed assemblies enabled the determination of physical linkages between resistance genes. On the other hand, NCBI761 was not accompanied by clinical data, excluding it from genotype-to-phenotype association studies.

1.2.2 Genes contributing to resistance in USWest352

Antibiotic resistance genes in the USWest352 and NCBI761 were identified using a custom annotation pipeline, based on BLAST searches against the ResFinder sequence databases [31] (see Methods).

A numerical summary of the resistance genes identified in USWest352 can be found in **Table 1.3**. In USWest352, our annotation pipeline identified 34 different resistance genes, with genes encoding AMEs (14 genes) or β-lactamases (14 genes) accounting for 99% of the 855 complete open reading frame (ORF) hits identified.

Table 1.2: Antibiotics considered in this study.

Antibiotic	Abbreviation	Class
Ampicillin	AMP	3rd-gen penicillin (β-lactam)
Cefazolin	CFZ	1st-gen cephalosporin (β-lactam)
Ceftazidime	CAZ	3rd-gen cephalosporin (β-lactam)
Ceftriaxone	CRO	3rd-gen cephalosporin (β-lactam)
Cefepime	FEP	4th-gen cephalosporin (β-lactam)
Gentamicin	GEN	aminoglycoside
Tobramycin	TOB	aminoglycoside
Ciprofloxacin	CIP	2nd-gen fluoroquinolone
Levofloxacin	LVX	2nd-gen fluoroquinolone

To quantify the effect of these genes on aminoglycoside and β-lactam resistance in USWest352, we performed association studies using logistic regression. In microbial association studies, datasets containing a population structure that reflects the selection of the phenotype of interest, particularly if this phenotype is under strong selection, can produce a high number of false positives [35]. To identify causative associations to resistance with confidence, we used an approach inspired by recent microbial genome-wide association studies that takes population structure into account [36, 37]. We designed regressors that capture genetic variance corresponding to the population structure, using multidimensional scaling (MDS) (see Methods). Using these as the only regressors in our predictive model allowed us to estimate the contribution of population structure to resistance.

Results from our logistic regression analysis are summarized in **Table 1.4**. We measured the predictive accuracy of our model using the area under the curve (AUC) metric, which is a value between 0.5 and 1.0. An AUC of 0.5 denotes a predictive accuracy no better than random, and 1.0 denotes perfect accuracy. We took the mean AUC value across 5000 bootstrap iterations and all drugs belonging to a specific class. We looked at the correspondence between FQN resistance (CIP/LVX) and population structure, and found a high correspondence between the two (mean AUC: 0.87). Including four FQN resistance mutations that inhibit the binding of topoisomerase inhibitors [38, 39] increased predictive accuracy only moderately (mean AUC: 0.94), confirming that FQN resistance exhibits a strong alignment with population structure. In the case of aminoglycoside resistance, the correspondence with population structure was considerably lower (mean AUC: 0.63), consistent with a predominantly peripheral genome location. The correspondence between β-lactam resistance and population structure was higher, however, with a mean AUC of 0.79. This suggests that vertical transmission comprises a larger component of the total transmission of β-lactamases, perhaps because of integration of

Table 1.3: Genes known to confer resistance against aminoglycosides, β-lactams and fluoroquinolones found in 352 ExPEC genome assemblies (USWest352).

Aminoglycoside

Gene name	Gene/Enzyme class	Complete ORFs	Genomes containing gene
aph(3")-Ib	AME: O-Phosphotransferase	106	100
aadA5	AME: O-Adenyltransferase	99	99
aph(6)-Id	AME: O-Phosphotransferase	101	96
aac(6')-Ib-cr	AME: N-Acetyltransferase	62	62
aac(3)-IId	AME: N-Acetyltransferase	37	37
aac(3)-IIa	AME: N-Acetyltransferase	27	27
aadA24	AME: O-Adenyltransferase	18	14
aadA1	AME: O-Adenyltransferase	16	12
aph(3')-Ia	AME: O-Phosphotransferase	14	13
aadA2b	AME: O-Adenyltransferase	10	9
aadA1b	AME: O-Adenyltransferase	2	2
aadA16	AME: O-Adenyltransferase	2	2
ant(2")-Ia	AME: O-Adenyltransferase	2	2
aac(3)-VIa	AME: N-Acetyltransferase	1	1
rmtE	16S rRNA Methyltransferase	1	1

β-lactam

Gene name	Gene/Enzyme class	Complete ORFs	Genomes containing gene
blaTEM-1A	CTX-M β-lactamase	157	140
blaCTX-M-15	CTX-M β-lactamase	79	78
blaOXA-1	OXA β-lactamase	66	66
blaCTX-M-14b	CTX-M β-lactamase	18	18
blaCTX-M-27	CTX-M β-lactamase	7	7
blaCMY-2	AmpC β-lactamase	7	4
blaCTX-M-55	CTX-M β-lactamase	3	3
blaTEM-12	TEM β-lactamase	3	3
blaTEM-19	TEM β-lactamase	2	2
blaCTX-M-1	CTX-M β-lactamase	2	2
blaCTX-M-65	CTX-M β-lactamase	2	2
blaCTX-M-104	CTX-M β-lactamase	1	1
blaCARB-2	CARB β-lactamase	1	1
blaCARB-11	CARB β-lactamase	1	1

Fluoroquinolone

Gene name	Gene/Enzyme class	Complete ORFs	Genomes containing gene
qnrB6	PMQR	2	2
qnrB19	PMQR	2	2
qnrS2	PMQR	2	2
qepA1	PMQR	1	1
qepA4	PMQR	1	1

β-lactamases into the chromosome.

Adding the set of resistance markers identified by our annotation pipeline (a total of 34 genes and 4 point mutations) as regressors to the model substantially increased the accuracy for aminoglycosides (mean AUC: 0.92) and for β-lactams (mean AUC: 0.91). For the 3rd/4th-generation cephalosporins, which have been introduced in the clinic more recently than AMP and CFZ, the predictive accuracy was even higher (mean AUC: 0.96). These observations suggest that the resistance genes identified by our annotation pipeline contribute significantly to aminoglycoside and β-lactam resistance.

We identified three genes as major contributors to resistance against the aminoglycosides based on the magnitude of the decrease in predictive accuracy following their removal from the model. The three genes are *aac(3)-IIa* (GEN resistance), *aac(3)-IId* (GEN/TOB resistance), and *aac(6')-Ib-cr* (GEN/TOB resistance). A high correspondence between resistance and genome-based predictions including these three genes has been previously reported in *E. coli* and *K. pneumoniae* [29] and in porcine Enterobacteriaceae [31], although these earlier studies did not control for population structure.

We performed conjugation assays to show that these genes are sufficient to confer resistance in a naïve genetic background. Our experimental results were largely consistent with the conclusions of our logistic regression analysis. The exception was the association between the *aac(3)-II* genes and TOB resistance, where our regression analysis showed a causal relationship, while the conjugation experiments failed to transfer resistance along with the gene. This discordance may result from differences in specific genetic background. Also, it has been reported previously that the AAC(3)-II enzymes only confer intermediate resistance against TOB in ST131 strains [40]. In the case of *aac(3)-IIa*, our regression analysis also failed to predict a causal relationship because *aac(3)-IIa* was always in the presence of another TOB resistance gene, *aac(6')-Ib-cr*.

We identified six genes as major contributors to resistance against 3rd/4th-generation cephalosporins: *blaCTX-M-14b*, *blaCTX-M-15*, *blaCTX-M-27*, *blaCTX-M-55*, *blaTEM-12* and *blaCMY-2*. These are all known to confer ESBL resistance [41].

Table 1.4: Contribution of population structure to resistance and predictive value of 38 resistance markers.

Antibiotic	mean AUC (permuted response variable)	mean AUC (population structure regressors only)	mean AUC (population structure and resistance markers)
Gentamicin	0.49	0.68	0.92
Tobramycin	0.50	0.59	0.92
Ampicillin	0.50	0.78	0.84
Cefazolin	0.50	0.77	0.84
Ceftazidime	0.50	0.82	0.96
Ceftriaxone	0.50	0.73	0.93
Cefepime	0.50	0.84	0.97
Ciprofloxacin	0.50	0.87	0.94
Levofloxacin	0.51	0.87	0.94

1.2.3 Exploratory factor analysis reveals two adaptive pathways of multidrug resistance

In order to identify global patterns in the distribution of acquired resistance genes identified by our annotation pipeline in USWest352, we performed exploratory factor analysis. This analysis combined multidimensional scaling (MDS) with unsupervised classification to produce a nondeterministic clustering of genomes based on their resistance gene composition. Visualization of these clusters along principal components representing the largest explained variance of gene composition allowed identification of the most successful adaptive solutions (see Methods).

We generated an input distance matrix for this analysis that contained measures of dissimilarity between each pairing of genomes' resistance gene composition. We then projected the matrix onto three principal components (PC1, PC2 and PC3). Sample genomes containing an identical set of resistance genes were collapsed (i.e. one representative was chosen), leaving 87 genomes with unique combinations of resistance genes, each combination representing a different evolutionary outcome. In effect, this minimizes the impact of clonal expansion in the analysis. For increased accuracy, the relative positions of genomes in the final output were calculated based on how their projected positions clustered over 3418 bootstrap replicates (see Methods). The distribution of gene combinations along PC1, PC2 and PC3 is shown in **Figure 1.1**.

Variance along PC1 was attributable to the presence of two APH genes that are known to confer STR resistance, *aph(3")-Ib* and *aph(6)-Id* (**Figure 1.1 a**). We found that 94% of the 87 genomes had one of two configurations, containing neither or both genes. Inspection of the assembled contigs containing these two genes showed that this high association between *aph(3")-Ib* and *aph(6)-Id* resulted from close physical linkage, occurring inside an IS240-type insertion element and with

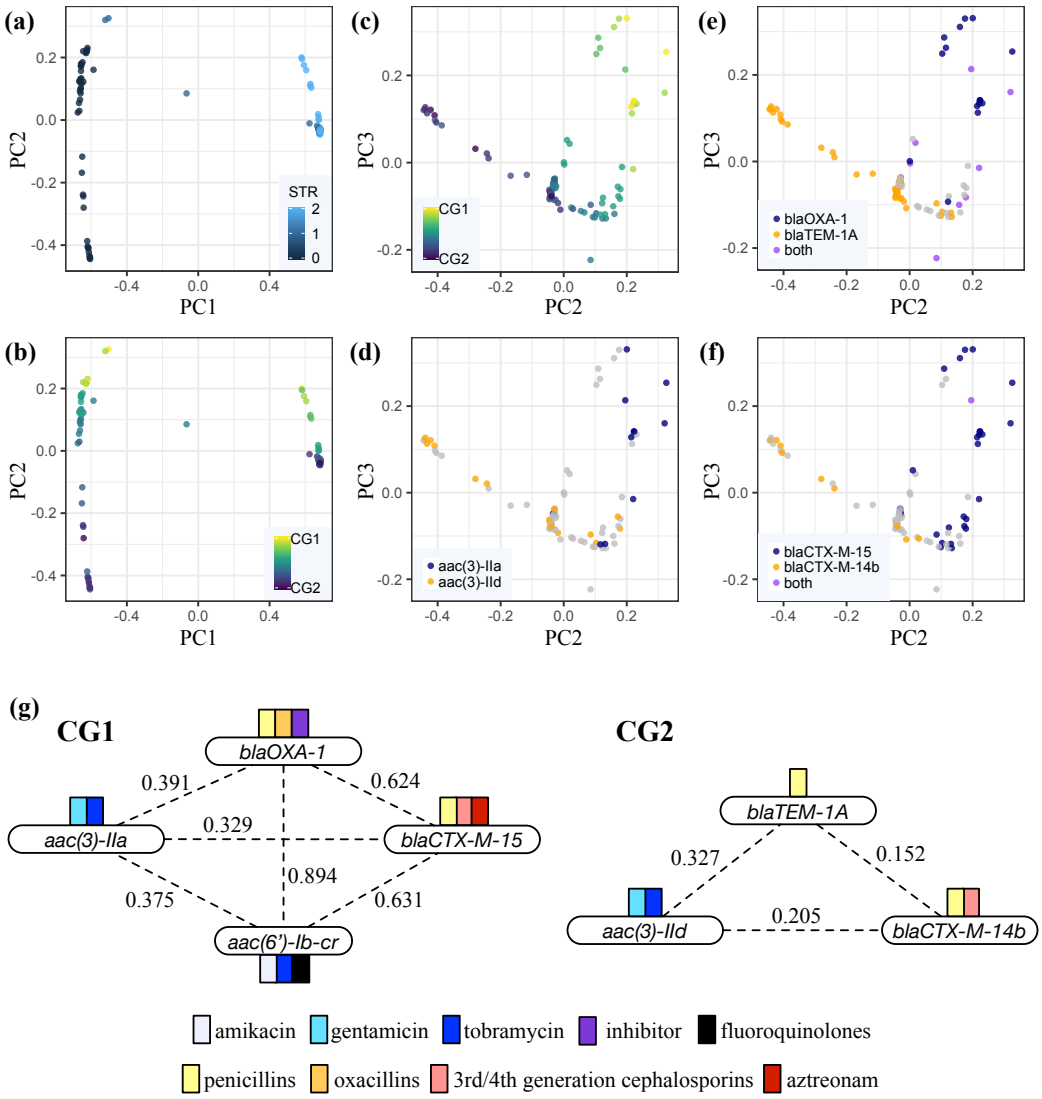

Figure 1.1: Clustering of genomes based on their composition of acquired antibiotic resistance genes. (a-f) MDS plots show the relatedness of 352 ExPEC genomes based on their composition of acquired resistance genes. The clustering of genomes is shown across three principal components PC1, PC2 and PC3: **(a,b)** across PC1 and PC2, **(c-f)** across PC2 and PC3. Colors indicate the presence of specific resistance genes, chosen based on their frequency of occurrence and distribution pattern: **(a)** streptomycin resistance genes, (2) both *aph(3")-Ib* and *aph(6)-Id*, (1) one of *aph(3")-Ib* or *aph(6)-Id*, (0) neither; **(b,c)** relative frequency of CG1 genes and CG2 genes; **(d-f)** mutual exclusion between specific CG1 (navy) and CG2 (orange) gene pairs with overlapping resistance profiles; in some cases both genes are present (purple) or neither (grey). **(g)** Two highly successful gene combinations, CG1 and CG2, represent adaptive solutions evolved in the presence of selective pressures by multiple antibiotics. The degrees of association (Jaccard index) between pairs of genes are provided along dotted lines.

overlapping ORFs. These two genes likely encode a single transcriptional unit, and are part of a STR resistance operon, based on previous reports [42]. Given the high representation of this two-gene combination (in 40% of 87 combinations), it was not

surprising that their presence or absence comprised the largest portion of variance in the USWest352. On the other hand, *aadA5*, which also occurred frequently (in 45% of 87 combinations) did not cluster with *aph(3")-Ib* and *aph(6)-Id* (data not shown). Together, these two results support the idea that the MDS projection along PC1 captured a highly prevalent gene interaction (specifically linkage between two genes) as opposed to an arbitrary representation of variance weighted solely by the gene frequency, validating our approach.

PC2 and PC3 resolved two gene groups to which genomes could be assigned based on the presence of seven resistance genes (**Figure 1.1 b** and **Figure 1.1 c**). Genomes assigned to the first group included some combination of *aac(3)-IIa*, *aac(6')-Ib-cr*, *blaOXA-1* and *blaCTX-M-15*, while genomes assigned to the second group included some combination of *aac(3)-IId*, *blaTEM-1A* and *blaCTX-M-14b*. Figure 1g shows the degree of co-occurrence of gene pairs in CG1 and CG2, using the Jaccard index (JI) [43]. Of the 87 unique combinations, 62% could be assigned unambiguously to one of the two groups, while only 15% did not contain any of these seven genes, so these two groups explain a large portion of the variance as well.

Looking at resistance profiles, we found that the combined resistance profile for each group of genes represents a largely complementary set of phenotypes (**Figure 1.1 g**), with individual resistance genes within each group being largely non-redundant in function. We therefore named these two groups Complementarity Groups 1 and 2 (CG1, CG2). Comparing CG1 and CG2, we also noted that each group exhibits a combined resistance profile that overlaps for six antibiotics frequently used in the clinic that have a strong signal in our logistic regression: AMP, CAZ, CRO, FEP, GEN, and TOB. This suggests that the two groups represent two distinct adaptive solutions driven by overlapping resistance profiles.

1.2.4 Pervasive antagonisms drive the segregation of CG1 and CG2

Assuming a similar selection, the evolution of parallel trajectories means that genes with similar resistance profiles are mostly found in separated clusters across PC2-PC3 space (**Figure 1.1 d-f**). The two strongest examples are *aac(3)-IIa* and *aac(3)-IId*, which provided resistance against the same antibiotics (GEN/TOB), and never co-occurred (**Figure 1.1 d**). The genes *blaCTX-M-14b* and *blaCTX-M-15* provided a second example, both of which confer ESBL resistance and occurred together only in 2% of the 87 combinations (**Figure 1.1 f**). We hypothesized that the striking segregation across these two groups of genes may be driven by mutual

antagonism arising from functional overlaps between groups. Meanwhile we also hypothesized that close physical linkage would help preserve associations between genes of the same complementarity group.

1.2.5 Phylogenetic distribution of CG1 and CG2 genes

We generated a phylogenetic distribution of gene combinations corresponding to CG1 and CG2 using a neighbor-joining tree based on 537,420 polymorphic sites across the genomes of USWest352 (**Figure 1.2**).

Gene combinations representing CG1 and CG2 were widely distributed across MLSTs (10 and 14 different MLSTs, respectively). Within individual MLSTs, CG gene representation was highly heterogeneous. For instance, in ST131, the numbers of CG1 and CG2 genes occurring in a genome ranged from 0 to 4 and 0 to 2, respectively and in ST648, CG2 is found as the full three-gene combination and in two different two-gene configurations (*blaTEM-1A* with *aac(3)-IId*, and *blaTEM-1A* with *blaCTX-M-14b*). Taken together, the weak linkages between some of the pairs, the ubiquitous distribution of CG1 and CG2 genes across MLSTs, and the overall heterogeneity within individual MLSTs, highlight the large degree of genomic plasticity involved in the evolution of MDR. This pattern of pervasive horizontal transfer and recombination is consistent with previous studies of genomic variation in prokaryotes [16, 17].

The high variability of CG1 and CG2 representation within individual MLSTs (including epidemic ones such as ST131) also implies that the success of epidemic strains is unlikely to be primarily driven by the acquisition of a particular combination of drug resistance genes. Indeed, in the UW233 dataset, which represents the unbiased sampling of the clinical strains isolated in a hospital, the majority of clonally expanded MLSTs (e.g., ST73, ST95, ST127) remained susceptible to most antibiotics. Further supporting this conclusion, previous literature showing that for a number of pathogens including ExPEC, only a fraction of strains with identical drug resistance phenotypes are successful, leading to the hypothesis that additional selective advantages are likely involved [44, 45].

CG1 and CG2 genes appeared to be highly concentrated in certain MLSTs across the phylogeny (e.g., ST10, ST44, ST131, ST405 and ST648), and in some cases appeared to be expanded uniformly in smaller monophyletic clades (**Figure 1.2**). These observations suggest that clonal expansion contributes to the dissemination of drug resistance to some degree, as previously reported for *blaCTX-M-15*

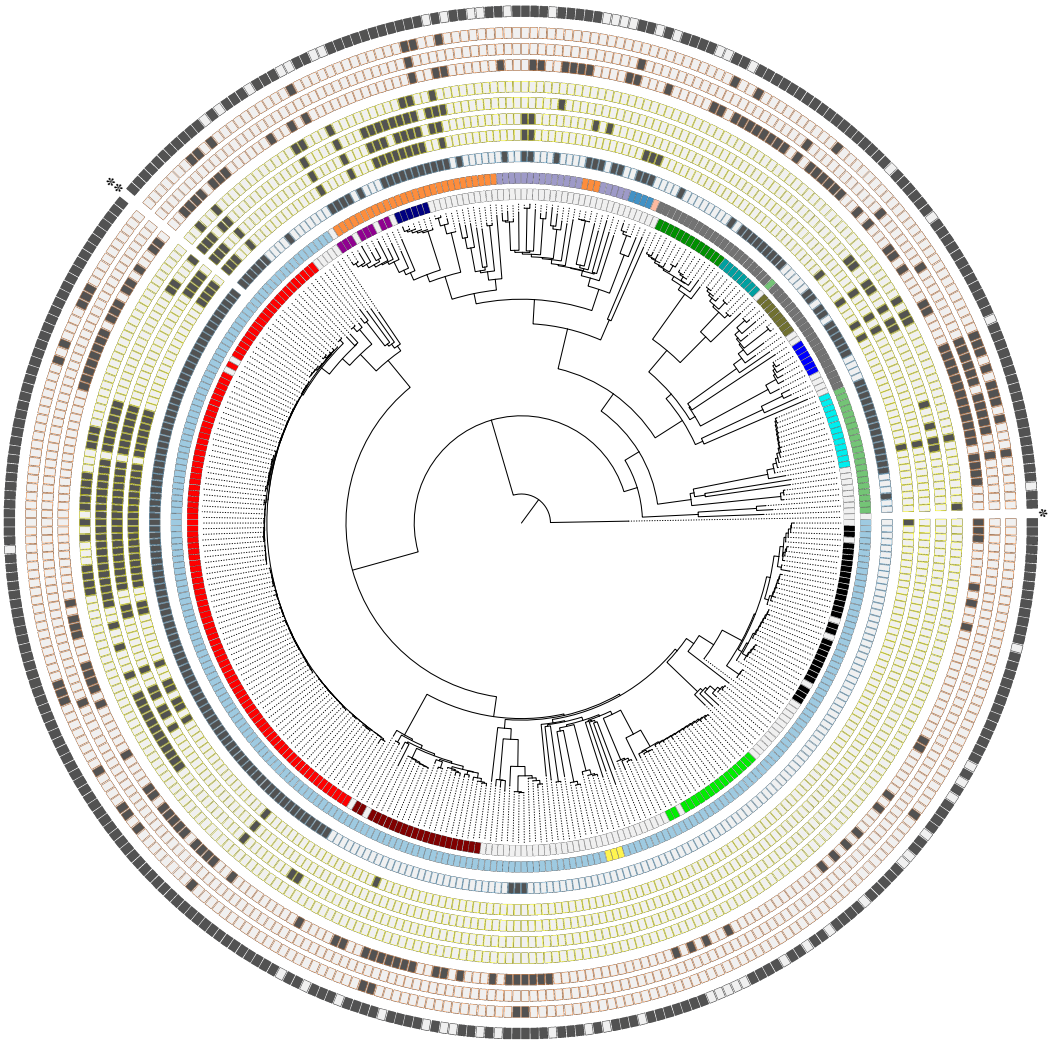

Figure 1.2: Phylogeny of 352 ExPEC isolates obtained from two hospitals on the U.S. West Coast. Phylogenetic relationships among 352 isolates are represented as a neighbor-joining tree, computed based on pairwise comparisons of 537,420 genomic SNPs. The Circos diagram surrounding the phylogeny indicates, for each phylogenetic taxa, its MLST and phylotype classifications, the presence of selected resistance markers and the presence of an IncF plasmid replicon. Circular panels, from the innermost to outermost, represent the following. Panel 1, MLST: ST10 (purple), ST38 (olive), ST44, (navy), ST69 (green), ST73 (maroon), ST95 (black), ST127 (lime), ST131 (red), ST393 (teal), ST405 (blue), ST648 (cyan). Panel 2, phylotype: A (orange), B1 (purple), B2 (light blue), C (dark blue), D (dark grey), E (light red), F (green), U (yellow). Panel 3 (outlined in blue), fluoroquinolone resistance mutation GyrA-S83X. Panels 4-8 (outlined in yellow), CG1 genes: *blaOXA-1*, *aac(6')-Ib-cr*, *blaCTX-M-15*, and *aac(3)-IIa*, respectively. Panels 9-11 (outlined in orange), CG2 genes: *blaTEM-1A*, *blaCTX-M-14b* and *aac(3)-IId*, respectively. Panel 12 (outlined in grey), the presence of any IncF replicon. *Outgroup: *Klebsiella pneumoniae* **Reference genome: E. coli EC958

in ST131 and ST405 [30, 46, 47]. Note that FQN resistance mutations in *gyrA* are often associated with these expansions, a pattern that appears more generally across MDR pathogens [45]

Despite the contribution of clonal expansion, the high genomic plasticity suggests the divergence of sublineages bearing CG1 and CG2 genes by clonal expansion should only make limited contributions to the observed segregation between CG1 and CG2 genes (**Figure 1.1** and **Figure 1.2**). This is more so, because these strains come from only two geographically proximal hospitals, and thus the sublineages likely had access to each other. Indeed, we find instances of MLSTs where both CGs are present (ST10, ST131, ST617 and ST648), showing that the strains had access to both evolutionary solutions.

Regional differences in the distribution of CG1 versus CG2-bearing samples between the two hospitals are not major contributors to the observed segregation either, since the segregation is present even when samples from each hospital are considered independently (not shown).

1.2.6 Determining the co-selection landscape of USWest352 using a random distribution model

The statistical significance of our observations was established using a random distribution model. We reasoned that, given enough gene flow, for two given resistance genes the probability of co-occurrence should be determined by their individual empirical frequencies in the dataset. In this model, we interpreted overrepresentation as indicative of positive co-selection, and underrepresentation as evidence of mutual antagonism.

In our model we included the ten most frequent resistance genes: *aadA5*, *aph(3")-Ib*, *aph(6)-Id*, and all seven CG1 and CG2 genes. We measured co-occurrence between each pairing using the Jaccard index (JI) [43] and detected significant associations by permutation test. To adjust for multiple comparisons, we used a significance level α of 0.05 (see Methods).

To see whether geographically proximal strains have access to both solutions, and to confirm that when both solutions co-exist, they remain mutually antagonistic, we represented significant associations for USWest352 (**Figure 1.3 a**). To minimize the effect of population structure, we also restricted our analysis to strains belonging to a homogenous phylogenetic context: strain ST131 (n=161, **Figure 1.3 b**). Finally, to see whether these observations could be generalized to a variety of regional contexts, we extended our analysis to the NCBI761 dataset (**Figure 1.3 c**).

Overall, our permutation tests supported the presence of extensive co-selection between gene pairs within CG1 and CG2 and of mutual antagonism across

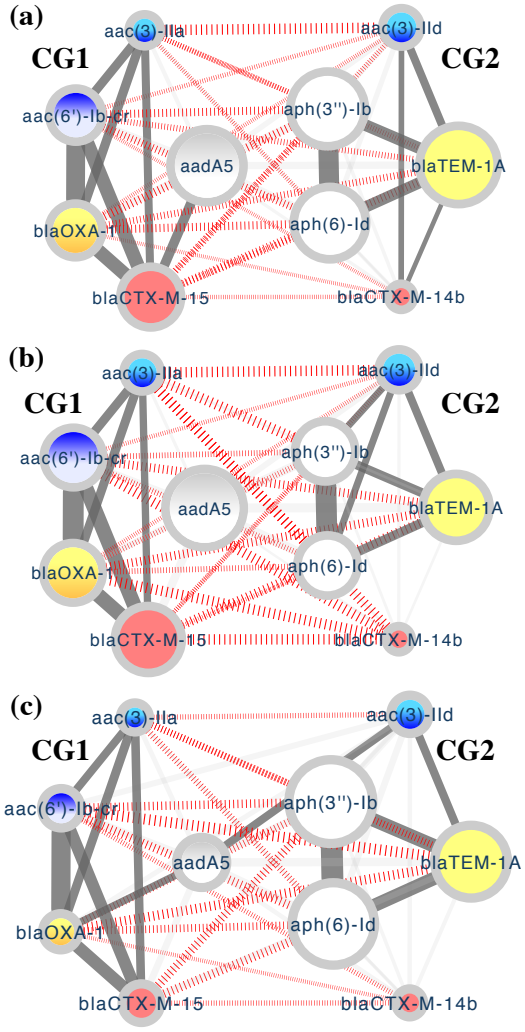

Figure 1.3: Positive and negative co-selection predicted between CG1 and CG2 genes. Pairwise associations between the ten most frequently occurring acquired resistance genes in the USWest352 dataset, predicted against a random distribution model for the following datasets: **(a)** USWest352, **(b)** 161 genomes belonging to the ST131 (gathered from USWest352 and NCBI761), and **(c)** NCBI761. Significant positive associations are indicated by dark grey links between nodes, while significant negative associations are indicated by red hashed links. Non-significant associations are shown as light grey links. Significance was determined by controlling for the false discovery rate at a significance level α of 0.05. Nodes representing genes are colored to represent their resistance profile: GEN (light blue), TOB (blue), amikacin (alice blue), penicillins (yellow), oxacillins (orange), 3rd/4th-generation cephalosporins (salmon), STR (white), spectinomycin (grey). Full resistance profiles for CG1 and CG2 genes are provided in **figure 1g**.

gene pairs in different subnetworks. This was true for a single, epidemic strain (ST131) and in a variety of epidemiological and geographical contexts. Additionally, in CG2 we found strong support for co-selection between *blaTEM-1A* and the STR resistance genes *aph(3")-Ib* and *aph(6)-Id*. On the other hand, we found that positive association patterns for *aadA5* were more inconsistent.

1.2.7 Mutual antagonism between resistance genes with similar functions

To confirm a possible connection between mutual antagonism and functional redundancy, we determined the degree of association within groups of genes that have similar resistance profiles. We assigned 108 different resistance genes identified in NCBI761 to 13 functional groups and measured underrepresentation of gene co-occurrence within each functional group against our random distribution model. Gene-to-group assignments for specific genes are provided in **Table 1.5**, and rationale for their assignments are provided in the Methods.

P-values indicating significance for underrepresentation of within-group co-occurrence are shown in Table 5. Gene co-occurrence was underrepresented in 8 groups: Groups 1, 2, 5 (AMEs), 7, 8, 10, 12 (β-lactamases), and 13 (PMQRs), with significance assessed at 95% over 1000 genome permutations. Group 3 had a p-value close to the 95% threshold (94.6%). Two groups of genes (Groups 6 and 9) showed non-significant p-values despite not having any instances of within-group co-occurrence, but this was likely due to insufficient sample size. Other exceptions were Groups 4 and 11, both of which include two genes co-mobilized in insertion elements. In the case of Group 4, the two genes are *aph(3")-Ib* and *aph(6)-Id*. Being part of an operon, overrepresentation of their co-occurrence against random distribution was expected. In Group 11, most of the observed co-occurrences (6 of 7) involved *blaOXA-1* and other OXA-type genes, some of which (4 out of 7) confer resistance to carbapenems, which could explain their co-occurrence with *blaOXA-1*. In addition, *blaOXA-1* was physically linked to *aac(6')-Ib-cr* inside an IS240-type insertion element, and thus it could have been passively co-selected through a *aac(6')-Ib-cr*-driven selection. Finally, by generating permutations of the gene-to-group assignments we confirmed that our significant p-values were not skewed by group size or attributable to multiple comparisons (grand-p-value=0.000; see Methods).

Overall, our random distribution model confirmed widespread mutual antagonism between genes in the same functional group when sample sizes were sufficient. The only exceptions we found correspond to two resistance genes that are very tightly linked and co-mobilized in an MGE, or when two genes shared an operon, potentially requiring both units for high level resistance.

Table 1.5: Negative co-selection within functional groups.

Group	Class grouping	Functional grouping	Genes	Sample Size	P-value
1	AAC(3)	+GEN, +TOB, -AMK	*aac(3)-IIa*, *aac(3)-IId*, *aac(3)-IV*[1], *aac(3)-Ia*, *aac(3)-VIa*, *ant(2")-Ia*[2]	112	0.001
2	AAC(6')	+TOB, +AMK, -GEN	*aac(6')-33*, *aac(6')-Ian*, *aac(6')-Ib*, *aac(6')-Ib-cr*, *aac(6')-Ib3*, *aac(6')-Il*	74	0.010
3	ANT	+STR, +SPC	*aadA1*, *aadA13*, *aadA16*, *aadA2*, *aadA22*, *aadA23*, *aadA24*, *aadA2b*, *aadA3*, *aadA5*	255	0.054
4	APH(3"), APH(6)	+STR, -SPC	*aph(3")-Ib*, *aph(6)-Id*	365	1.000
5	APH(3')	+KAN, +NEO, +PRM	*aph(3')-IIa*, *aph(3')-Ia*, *aph(3')-VI*, *aph(3')-VIa*, *aph(3')-VIb*	88	0.007
6	16S RMT	+GEN, +TOB, +AMK	*armA*, *rmtB* and *rmtC*	19	0.533
7	CMY (Class C)	+carbapenems +ESBL	*blaCMY-111*, *blaCMY-16*, *blaCMY-2*, *blaCMY-23*, *blaCMY-34* *blaCMY-4*, *blaCMY-42*, *blaCMY-44*, *blaCMY-new*, *blaCMY-6*	52	0.010
8	CTX-M (Class A)	+ESBL	*blaCTX-M-123*, *blaCTX-M-14b*, *blaCTX-M-15*, *blaCTX-M-199*, *blaCTX-M-2*, *blaCTX-M-24* *blaCTX-M-27*, *blaCTX-M-3*, *blaCTX-M-55*, *blaCTX-M-64*, *blaCTX-M-65*	135	0.000
9	KPC (Class A)	+carbapenems	*blaKPC-2*, *blaKPC-3*[3], *blaKPC-4*	29	0.245
10	NDM (Class B)	+carbapenems +all β-lactam, except ATM	*blaNDM-1*, *blaNDM-21*, *blaNDM-4*, *blaNDM-5*, *blaNDM-6*, *blaNDM-7*, *blaNDM-9*	61	0.001
11	OXA (Class D)	+oxacillins some carbapenemases[4]	*blaOXA-1*, *blaOXA-10*, *blaOXA-163*, *blaOXA-181*, *blaOXA-2*, *blaOXA-4*, *blaOXA-48*, *blaOXA-9*	77	0.296
12	TEM (Class A)	+penicillins only (99% of group)	*blaTEM-116*, *blaTEM-135*, *blaTEM-156*, *blaTEM-176*, *blaTEM-1A*, *blaTEM-20* *blaTEM-210*, *blaTEM-215*, *blaTEM-26*, *blaTEM-30*, *blaTEM-32*, *blaTEM-57*	200	0.000
13	PMQR	+CIP, +LVX	*qepA1*, *qepA4*, *qnrA1*, *qnrB10*, *qnrB4*, *qnrB52*, *qnrB6*, *qnrB9*, *qnrE1*, *qnrS1*, *qnrS2*, *qnrVC4*	54	0.044

[1] Has broad substrate specificity range that includes GEN, TOB, AMK and apramycin.

[2] Does not encode an AAC(3) enzyme, but has similar resistance profile, *e.g.*, confers resistance against GEN and TOB, but not AMK.

[3] Reported to confer resistance against cefoxitin, a cephamycin.

[4] *blaOXA-48*, *blaOXA-163* and *blaOXA-181* additionally confer carbapenem resistance.

1.2.8 CG1 and CG2 genes can be found together in conjugative plasmids

Co-selection is facilitated by close physical association between the relevant pairs of genes. We checked this hypothesis experimentally by conjugation of strains bearing CG1 and CG2 genes. For CG1, in one USWest352 isolate we mapped all four CG1 genes to an IncF, conjugation-competent plasmid, confirming their physical linkage in this instance. We then used a primer-walking strategy to determine the relative proximity and orientation of the four genes in several USWest352 isolates. In two isolates (U90, U95), we found an IS240-type insertion element containing *aac(6')-Ib-cr* and *blaOXA-1* located approximately 1.2 kb upstream of *aac(3)-IIa*, and in three isolates (U2, U46, U90) *blaCTX-M-15* was nearly adjacent and upstream of *aac(6')-Ib-cr*. For CG2, we mapped all three CG2 genes to a conjugative plasmid in one isolate, again showing that all the genes can be physically linked in the same extrachromosomal element.

1.2.9 Physical linkage map of acquired antibiotic resistance for NCBI761

We used complete genomic assemblies of NCBI761 to verify the associations of co-selection detected in our random distribution model as physical linkages. These complete assemblies also allowed the identification of other gene associations including genes for which we don't have phenotypic data.

The degree of physical linkage between gene pairs where each constituent gene is present more than ten times in NCBI761 was estimated using JI (**Figure 1.4**). Overall, the physical linkage profile for CG1 and CG2 gene pairs is highly consistent with the degree of positive co-selection predicted using our random distribution model.

We detected three pairs of genes with very tight physical linkage, only one of which is specific to CG1: *aac(6')-Ib-cr* and *blaOXA-1* (JI: 0.82). The second pair is *aph(3")-Ib* and *aph(6)-Id* (JI: 0.95), which we found to be more closely associated with CG2 than CG1. These two genes, which have been reported to encode a single transcriptional unit as part of a STR resistance operon [42], are co-oriented with overlapping ORFs and could together provide high-level STR resistance. Finally, the third gene pair exhibiting tight genetic linkage includes *oqxA* and *oqxB* (JI: 0.9), which encode two subunits of the RND family efflux pump often found as an operon [48].

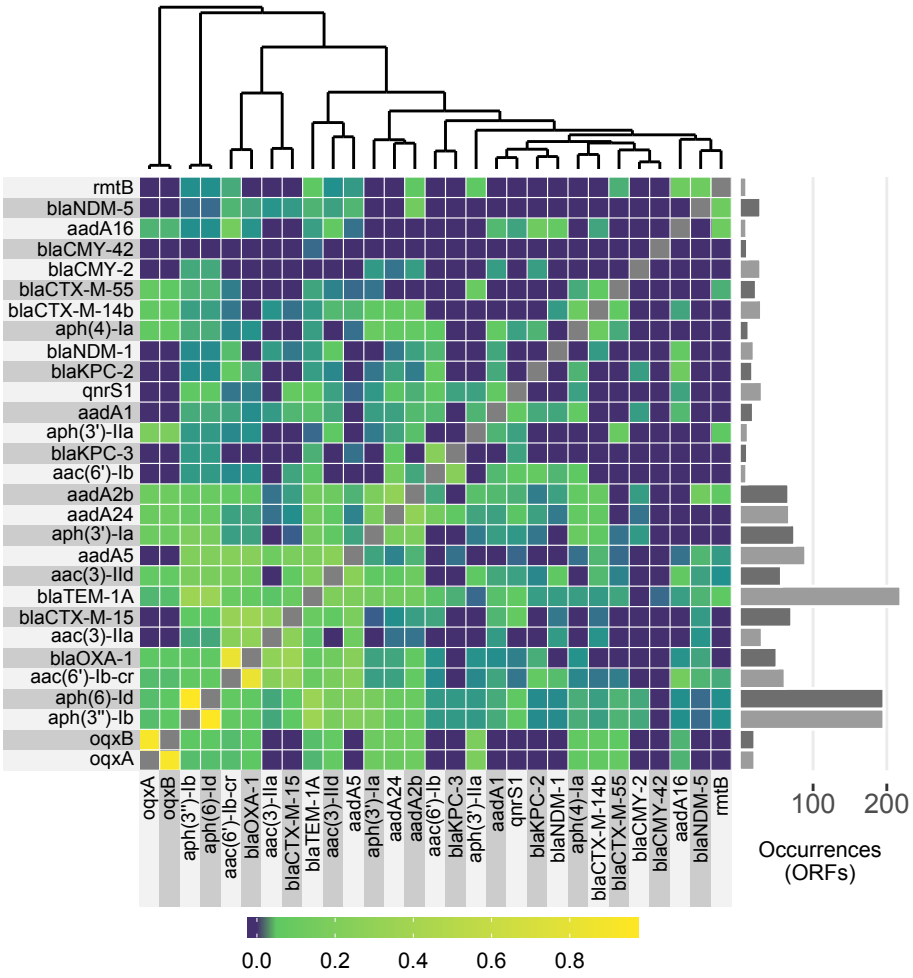

Figure 1.4: Physical linkage landscape of 761 *E. coli* genomes from NCBI. The heatmap represents a physical linkage landscape, or the degree of co-occurrence in the same plasmid or chromosome, between all pairings of aminoglycoside, β-lactam and fluoroquinolone resistance genes identified among 761 complete *E. coli* genomes from NCBI. Genes with ten or more occurrences across the dataset were included in the heatmap. The degree of co-occurrence between two genes was measured using the Jaccard index. The dendrogram above the heatmap, generated using kmeans, shows the clustering of genes according to their physical linkage profiles. The number of times each gene was found in the dataset is shown to the right. The legend below the heatmap shows the heatmap colors corresponding to Jaccard indices ranging from 0.0 to 1.0.

This analysis also detected another less frequent (n=17) gene combination whose gene composition is consistent with our definition of a complementarity group. We named this combination CG3; it includes *aph(4)-Ia*, which according to the literature confers *hygB* resistance [49], *aac(3)-IV*, which protects against GEN, TOB, and AMK [50] and *blaCTX-M-14*, which is an ESBL [41]. Within this CG, we see

again a clear functional diversification, as well as mutual antagonism with genes present in other CGs (**Figure 1.5**).

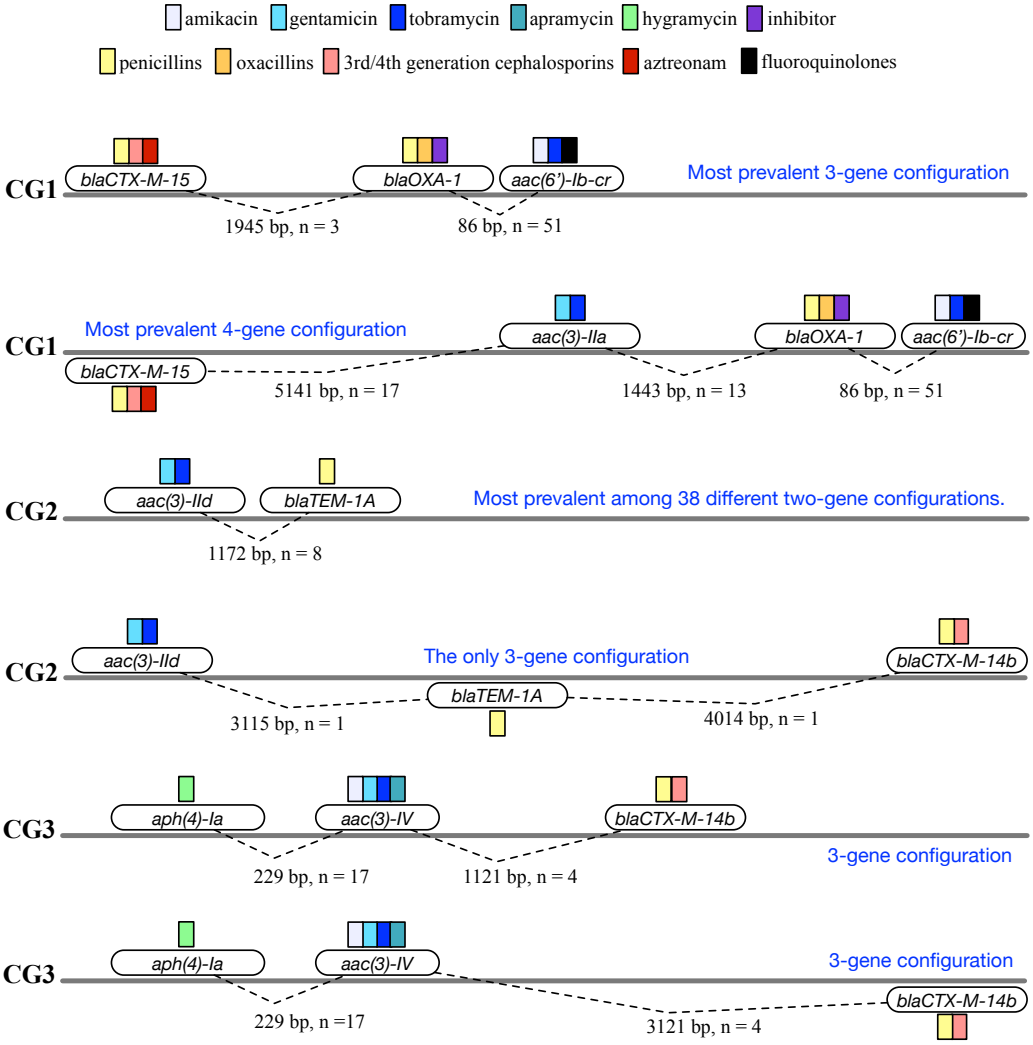

Figure 1.5: Configurations of CG1, CG2 and CG3. A few selected configurations for CG1, CG2 and CG3 are shown.

1.2.10 Multiple configurations between CG1 and CG2 gene pairs

MRRs represent complex adaptive solutions. CG1, CG2, and CG3 can be understood as successful adaptive solutions comprising the portions of MRRs that encode the set of resistances attributable to these respective CGs.

The order of genes and distances separating genes in CG1, CG2 and CG3 are mapped in **Figure 1.6 a**, for all occurrences in NCBI761. These CGs, and more specifically, any given set of resistance genes in the NCBI761, can be found in a plurality of configurations, as defined by gene order, proximity and co-strandedness.

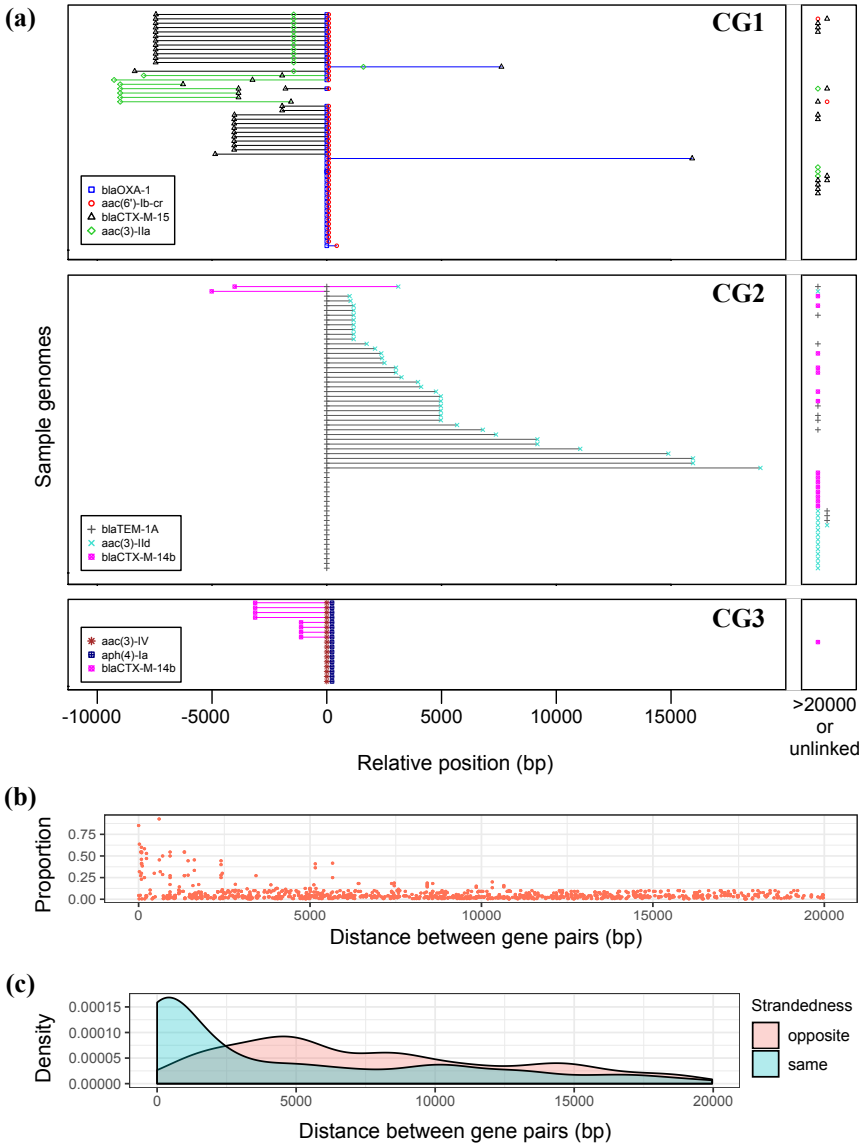

Figure 1.6: Physical linkage configurations for adaptive solutions: proximity and co-strandedness of genes. (a) Diagrams representing different configurations (order and proximity) of the genes comprising CG1, CG2 or CG3, found in NCBI761. For each sample genome represented across the y-axis, genes that occur within 20,000 bps of each other are plotted with connecting lines, while genes that are farther apart than 20,000 bps and unlinked genes that occur on different molecules are plotted on a separate panel to the right. **(b)** A plot comparing the success of different configurations (proximity) of the same pair of resistance genes. The Success Index (SI) represented by the y-axis indicates, for each configuration of a given pair of genes, its proportional representation in NCBI761. Gene pairs that are physically linked in at least 10 genomes in NCBI761 are plotted. **(c)** Density plotted versus distance between resistance genes: gene-pair configurations are grouped solely based on co-strandedness (occurring on same or opposite strands), and without considering specific genes. Gene pairs that occur on the same or opposite strands are represented on two separate density plots.

For instance, *blaTEM-1A* and *aac(3)-IId* of CG2 are found in 38 different

configurations with distances between the two genes ranging widely from 995 to 17,154 bp. Any two CG1 genes can also display considerable variation in proximity, and can occur on the same or opposite strands. Although much less prevalent, CG3 also displays multiple configurations, in the linkage of *blaCTX-M-14b* to the other two CG3 genes.

Notably, the associations between genes within CG1 and CG2 are preserved across large distances, and even when there is no physical linkage between the genes. However, there is a clear trend favoring stability of association for configurations placing genes in high proximity. Two patterns emerge, more generally, across all 108 resistance genes identified in NCBI761. If we define success as the number of times the configuration is seen in the dataset (as an indicator of representation in that population), configurations that place a given pair of genes in higher proximity tend to be more successful (**Figure 1.6 b**). Additionally when two genes are in high proximity, configurations that place them on the same strand tend to be more successful (**Figure 1.6 c**).

Taken together, these directional trends observed across many co-existing configurations of the three CGs and other gene combinations suggest a gradual evolution, or fine-tuning of adaptive solutions, towards high proximity and co-strandedness for acquired resistance genes under selection. Further, given our previous findings of mutual antagonism, this fine-tuning process appears to favor gene combinations without significant overlaps in resistance profiles.

1.2.11 CG features portray success, consistent with the genetic capitalism model

Based on the observations in this study, we propose three features of successful gene combinations found as part of CGs. (1. Diversification) The member genes tend not to overlap in function, conferring resistance against different antibiotics, thereby presumably decreasing the per-resistance fitness costs associated with transcription, translation and plasmid upkeep in a process analogous to the evolution of operons [51]. The antagonisms driving this diversification could result from enzymatic competition for the same substrate [reference in review], and from recombination between highly similar sequences [52]. (2. Co-transferability) The member genes tend to occur close to one another, thus increasing the probability of co-transfer during recombination. Genes found close together inside immediately flanking insertion sequences stood out as having the highest co-transferability, lead-

ing to high evolutionary stability. (3. Co-strandedness) Orientation of member genes on the same strand likely increases transcriptional efficiency, as it is frequently seen in operons [51]. Location in the leading strand would have the added advantage of decreasing the incidence of deleterious mutations by avoiding a collision between the transcription and the replication machineries [53].

These three features underlie the progressive growth of CGs, and are consistent with the positive feedback loop proposed by genetic capitalism [22].

1.2.12 Proposed roles of mutual antagonism and genetic linkage during evolution of MDR

The evolution of MDR appears to involve three elements: (1) positive selection by the relevant antibiotics; (2) positive co-selection between resistance genes, amplified by proximity in the genome; and (3) mutual antagonism across genes with overlapping functions, both within and across adaptive solutions in parallel trajectories.

We propose that mutual antagonism within adaptive solutions leads to functional diversification by disfavoring gene combinations that contain genes with redundant functions. Across adaptive solutions, we propose that mutual antagonism underlies the contingent evolution of discrete gene combinations i.e. it constrains future evolutionary trajectories, leading discrete adaptive solutions along separate pathways [27]. Evolutionary contingency was originally demonstrated in the context of genetic adaptation of model proteins [25,54–58], where specific antagonistic interactions between pairs of mutations have a strong "founder effect", leading to distinct evolutionary trajectories [25,26,55]. We propose that in MDR evolution, mutual antagonism drives alternate trajectories by disfavoring the acquisition of functionally related genes. This mechanism is supported by the observation that occasional instances of co-occurrence between genes with large overlaps in function seem to be transient. For instance, in USWest352, only one strain carried both *blaCTX-M-14b* and *blaCTX-M-15*, and another had two copies of *blaCTX-M-15*.

In protein evolution, strong antagonisms between adaptive mutations can arise from physical constraints imposed by protein structure, for example because two antagonistic mutations alter the active site in incompatible ways [25,26]. In the context of MDR evolution, by contrast, antagonism is inherently weaker because it arises indirectly between different resistance elements, possibly from increased fitness costs, competition for substrate [reference in review], or gene loss by recom-

bination [52]. Also in MDR evolution, when the functional overlap is incomplete, mutual antagonism can be partially offset through additional selections by antibiotics that do not effectively overlap as substrates. This can be seen in the case of *blaCTX-M-14b* and *blaCTX-M-55*, which co-occur more frequently than other *blaCTX-M* pairs (JI: 0.114), and which can be found physically linked unlike other *blaCTX-M* pairs. Both genes have ESBL activity but only *blaCTX-M-55* confers aztreonam resistance [59]. Another example is *blaCTX-M-15* and *blaOXA-1*; both are penicillinases, but *blaOXA-1* also confers oxacillinase activity and resistance to $\beta-$-lactamase inhibitors, while *blaCTX-M-15* confers ESBL resistance (JI: 0.624). On the aminoglycoside side, both *aac(3)-IIa* and *aac(6')-Ib-cr* share resistance to TOB, but *aac(3)-IIa* also confers GEN resistance, whereas *aac(6')-Ib-cr* also confers resistance to AMK and to FQNs.

Another difference between genetic adaptation in single proteins and MDR evolution is that genetic linkage and co-strandedness between pairs of genes represent additional evolutionary constraints. In MDR evolution, the stable acquisition of combinations of resistance genes depends on their co-transferability, which is a function of physical distance and/or of direct co-mobilization within MGEs (for instance, within the innermost set of insertion elements). This predicts a trend toward a reduction in the distance between the genes in each group, as combinations involving gene configurations closer to each other will be more efficiently transferred and thus more successful over time. This idea is additionally supported by the diversity of configurations observed for CG1, CG2, and CG3 (**Figure 1.6 a**) and by the observation that configurations placing the genes in high proximity and on the same strand tend to be more successful (**Figure 1.6 b** and **Figure 1.6 c**).

1.2.13 Proposed model for the evolution of hypothetical CG

In this work, we defined CGs as groups of functionally complementary genes conferring resistance to multiple classes of antibiotics that are consistently linked in the population. CGs can be included in MRRs representing more complex adaptive solutions.

Associations between acquired resistance genes can be understood as fluid linkages maintained via the selective pressures imposed by consistent antibiotic usage, and for which, the fluidity is driven by stochastic recombination events. **Figure 1.7** illustrates how a hypothetical CG could arise.

The initial generation of successful adaptive solutions is portrayed in **Fig-**

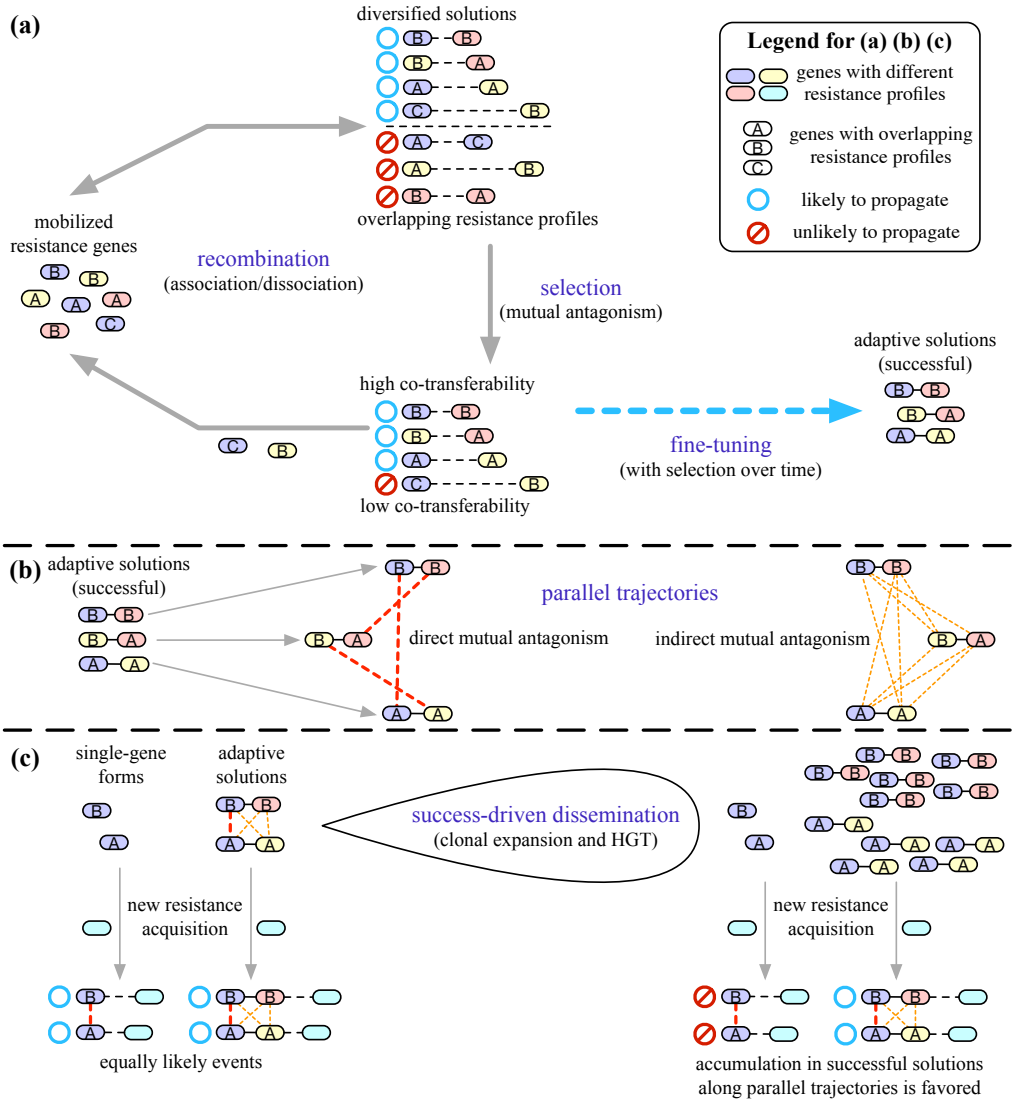

Figure 1.7: Model for evolution of acquired multidrug resistance. Our proposed model for the generation of hypothetical CGs is illustrated in three parts. **(a)** The generation of successful adaptive solutions characterized by high diversification, high co-transferability and co-strandedness (co-strandedness not shown). **(b)** The establishment of parallel evolutionary trajectories, and restriction of future additions by direct and indirect mutual antagonism across adaptive solutions. **(c)** The effects of success-driven dissemination of adaptive solutions: accumulation of additional resistance genes in successful adaptive solutions, and progression along parallel trajectories.

ure 1.7 a. The capture of individual resistance genes in MGEs represents a precursor to multidrug resistance. Diverse gene combinations arise from stochastic interactions between individual resistance genes. For a given pairing between two genes, the degree of stability is a function of proximity, which increases co-transferability and decreases the chance of becoming disassociated during future recombination events. Under selection, stability of that pairing is also determined by mutual antagonism

and this antagonism drives the functional diversification of the associations. An initial stable pairing represents an adaptive solution.

Over time, solutions that involve genes with less overlap in function, higher co-transferability, and a higher degree of co-strandedness will have a selective advantage and become the preferred configurations. Thus, there appears to be a process of fine-tuning of adaptive solutions by recombination and selection. This process is gradual, so at any given time we can see multiple intermediates within the population that may not exhibit all of these properties (putative intermediates for CG1, CG2 and CG3 shown in **Figure 1.6 a**).

The establishment of successful adaptive solutions limits future evolutionary trajectories because new gene acquisition is restricted by mutual antagonism with genes already part of the solution, based on functional overlap (direct mutual antagonism in **Figure 1.7 b**). Due to physical linkage, this mutual antagonism can also restrict the acquisition of new genes that do not have functional overlap (indirect mutual antagonism in **Figure 1.7 b**). Overall, the establishment of successful adaptive solutions leads to distinct evolutionary trajectories and decreases the genetic diversity available for future, more complex trajectories.

Finally, the ongoing dissemination of successful adaptive solutions by clonal expansion and HGT can further restrict evolutionary trajectories (**Figure 1.7 c**). The increased representation of successful solutions in the gene pool resulting from their dissemination facilitates the acquisition of additional genes by creating more opportunities for recombination with different genes (genetic capitalism). As a result, resistance genes tend to accumulate in successful solutions, favoring progression along existing parallel trajectories, while disfavoring both the realignment of existing trajectories and the emergence of alternate trajectories that are less successful.

The dissemination of successful adaptive solutions is not limited by geographical distance, as suggested by the widespread presence of CG1 and CG2 gene combinations across the two hospitals and across the various worldwide sampling locations represented in NCBI761 **Figure 1.8**. This is consistent with previous studies that showed geographical distances do not strongly influence the distribution of individual genes in the microbial gene pool [13, 60].

Given this generalized panmixia of antibiotic resistance genes and of adaptive solutions that we observe, it is still unclear why some successful ExPEC lineages (e.g., ST73, ST95, ST127) remain mostly susceptible to antibiotics frequently used in the clinic (e.g., GEN, TOB, extended-spectrum β-lactams), despite their geographic co-localization in the two US West Coast hospitals with a second set of

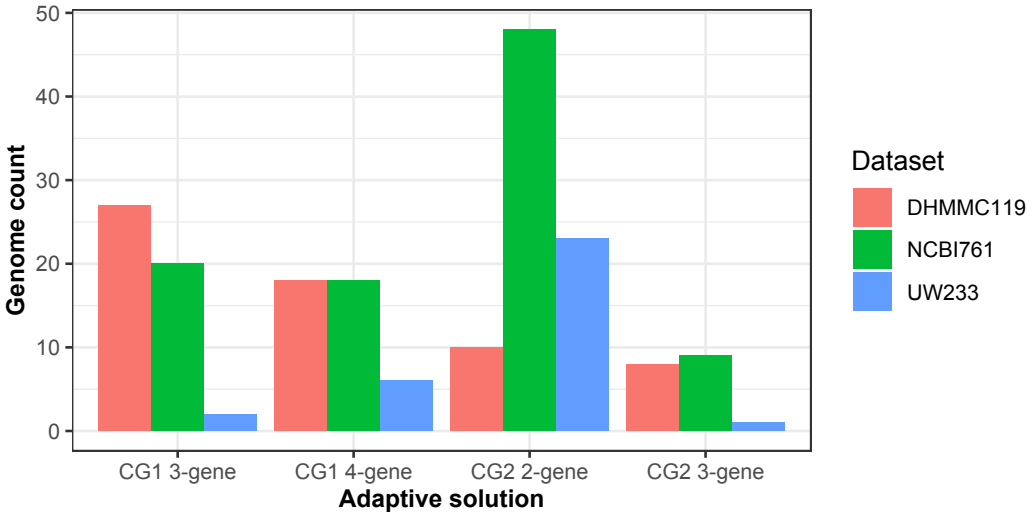

Figure 1.8: Number of genomes containing CG1 and CG2 gene combinations in DHMMC119, UW233 and NCBI761.

lineages (e.g., ST10, ST44, ST131, ST405 and ST648) that display a high degree of multidrug resistance. This partitioning of ExPEC lineages does not correlate with vertical transmission or with localization of infection (sepsis, UTI, etc.).

These two sets of lineages may represent niche specializations, as seen more generally in studies of spatial variation in microbial diversity of free-living microorganisms [61]. The acquisition of multidrug resistance via the accumulation of resistance genes in the peripheral genome could be seen as an adaptation of lineages that thrive in nosocomial environments. Indeed, there is some evidence that suggests three of these highly multidrug resistant ExPEC lineages (ST131, ST405 and ST648) are better suited for nosocomial environments; they are frequently implicated in nosocomial infections and often harbor a similar subset of acquired virulence factors [62, 63]. Susceptible lineages on the other hand could have large reservoirs in the environment; this would decrease their exposure to antibiotics relative to their total population.

1.2.14 Landscape of MDR association with plasmid replicons

To contextualize the evolution of CGs, and more generally the evolution of acquired MDR, from the perspective of plasmid evolution, we visualized the flow of acquired resistance genes within the gene pool in NCBI761 by generating a network representing physical linkages between resistance genes, and between resistance genes and plasmid replicons (**Figure 1.9 a**).

IncF replicons (IncFIA, IncFIB, IncFIC and IncFrepb) constituted central

nodes in the network (based on degree centrality metric), together comprising a large component of the total gene pool (44% of the 1802 ORFs corresponding to acquired resistance genes). IncF replicons showed a dense interconnectivity within the network, indicative of their frequent co-occurrence on the same plasmid. The combinatorial diversity of IncF plasmid replicons is well known and may help optimize replication for vertical versus horizontal transmission, and likely promotes compatibility among IncF plasmids [64, 65].

Replicons belonging to a variety of incompatibility groups showed considerable connectivity, including the following groups: IncAC, IncF, IncHI2, IncI1, IncK, IncN, and IncP. The presence of plasmids with mosaic replicons implies that the two parental (single replicon) plasmids would have had to be in the same host simultaneously long enough to allow recombination to occur [24]. Some groups stood out for being largely independent (IncBO, IncLM, oricolE), with no linkages to other replicon types, while still retaining linkages with resistance genes. The network also included a group of plasmids designated "Unknown", that did not contain identifiable replicons, and which were also largely independent. The presence of independent replicon types that tend not to participate in mosaicism suggests the presence of ecological or biological barriers.

Individual resistance genes tend to be linked to multiple replicons. There was a group of genes frequently linked to IncF replicons, but the degrees of linkage varied widely, ranging from 90% for *aadA5* to 27% for *blaCTX-M-14*, with the remaining being linked to non-IncF plasmids or integrated in the chromosome (**Figure 1.9 b**). This group of IncF-associated genes involved resistance to a variety of antibiotic classes, consistent with previous reports of completely sequenced IncF plasmids [65], and included all CG1 and CG2 genes. Other genes, however, are hardly associated with IncF replicons. Examples include OXA β-lactamases other than *blaOXA-1*, genes encoding the CMY β-lactamases, and NDM β-lactamases (with the exception of *blaNDM-4*).

In USWest352, we also found that CG1 and CG2 genes were largely linked with IncF plasmids, by monitoring the co-transfer of carbenicillin resistance and specific replicons following conjugation (see Methods). Among 68 conjugation-proficient plasmids bearing CG1 or CG2 genes, 97.3% had an IncF replicon. For the non-conjugative strains, we were unable to accurately ascribe CG1 or CG2 genes to specific replicons because our genomic assemblies were incomplete.

Overall, IncF plasmids, defined as plasmids containing at least one IncF replicon, appear to play a central role in the generation of adaptive solutions for

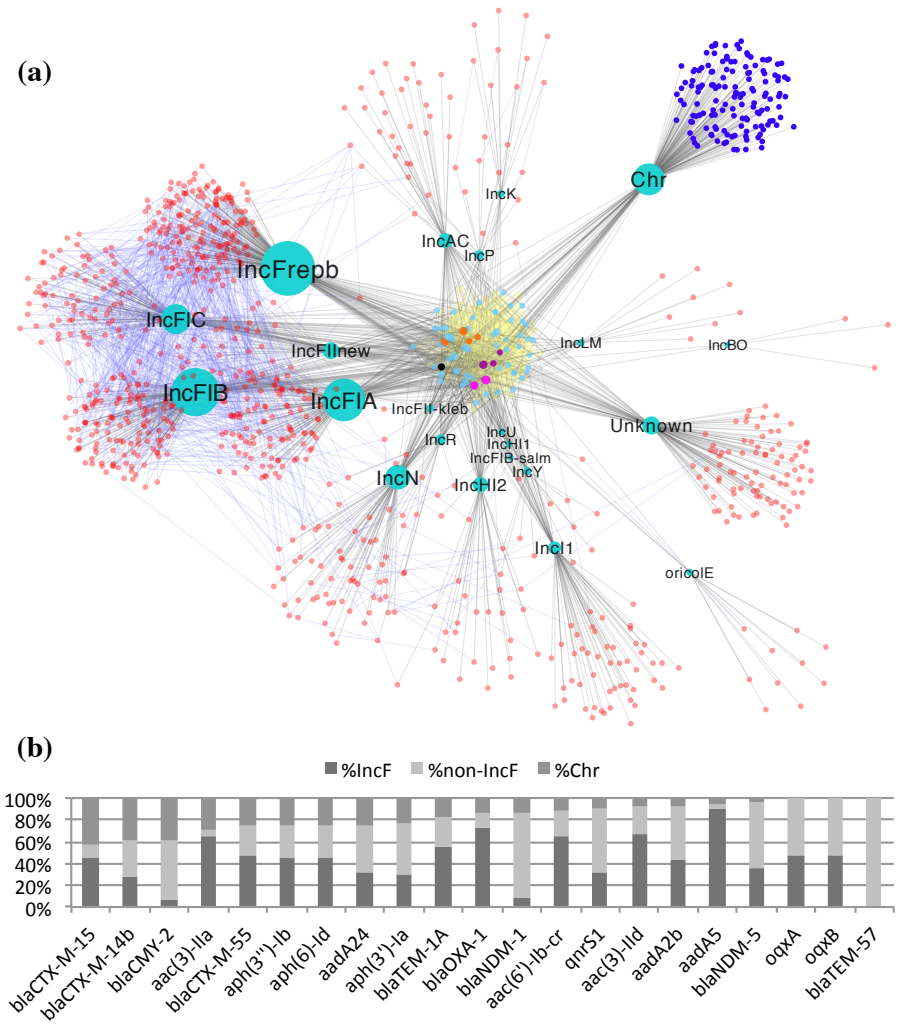

Figure 1.9: Genetic information flow network for plasmid replicons and acquired resistance genes. (a) The network depicts the complexity of gene flow across chromosomes and plasmids containing various replicon types, in the NCBI761 dataset. Nodes in the network represent genetic elements identified in NCBI761. In the periphery, plasmid replicons (small red nodes) and chromosomes (small blue nodes) are connected to larger nodes representing their classification (turquoise nodes). Acquired antibiotic resistance genes (light blue nodes) are positioned at the center of the network. Notable resistance genes are color-coded: CG1 (orange), CG2 (violet), *aph(3")-Ib* and *aph(6)-Id* (purple), *aadA5* (black). Physical linkages, or co-occurrence on the same molecule (chromosome or plasmid), are represented as edges, and occur between acquired resistance genes (yellow edges) and between plasmid replicons (blue edges), and between resistance genes and plasmid replicons (grey edges). Turquoise nodes that represent the chromosomes or various plasmid replicon types are sized based on the number of different resistance genes they are connected to i.e. degree centrality. In total, the network represent 1802 ORFs corresponding to aminoglycoside, β-lactam and fluoroquinolone resistance genes identified in NCBI761, which are distributed across 380 plasmids and 141 chromosomes. Node positions were estimated based on their interconnectivity to other nodes, using the edge-weighted, spring-embedded layout in Cytoscape. **(b)** Distribution of acquired resistance genes across chromosomes and plasmids, with plasmids divided into two categories, ones containing at least one IncF replicon, and ones without IncF replicons.

acquired MDR in *E. coli*. This is consistent with the distribution and properties of these plasmids (widespread distribution, conjugation competency) [44, 64], and is also in agreement with previous reports pointing to IncF plasmids as critical for the spread of resistance [64–66]. However, while central, the role of IncF plasmids as vehicles of evolution and spread of drug resistance in *E. coli* is not exclusive.

The complexity of linkages between replicons and between replicons and antibiotic resistance genes reinforces the idea that we are looking at intermediates in an adaptive process driven by selection involving random recombination events.

The presence of mosaic plasmids with replicons belonging to different incompatibility groups suggests they are the product of recombination between different parental plasmids. To our knowledge, widespread mosaicism involving IncF and non-IncF plasmids has not been reported before, with the exception of an IncF plasmid previously reported to have IncI1 and IncN replicons [67]. The redundancy in mechanisms of replication initiation that these plasmids have could broaden their compatibility range, potentially accelerating the process of genetic adaptation (MDR in this case). This was shown to be the case for IncF plasmids, where IncFII is free to diverge when associated with IncFIA or IncFIB because it does not participate in replication initiation when IncFIA or IncFIB is present. This drift can break down compatibility barriers, and increase the chance of compatibility with incoming IncF plasmids [64].

1.2.15 Capture of acquired resistance genes into the chromosome

In our analysis of NCBI761, we noted a large flow of acquired resistance genes in chromosomal locations (on average 18.0% of 1802 ORFs), although some genes were not found in the chromosome at all (**Figure 1.9 b**). Notably, three genes contributing to 3rd-generation cephalosporin resistance including *blaCTX-M-14b*, *blaCTX-M-15* and *blaCMY-2* were present in the chromosome in higher proportions (38.7% to 42.5%) than any other resistance gene. This was true despite low degrees of physical linkages between these three genes (**Figure 1.4**), indicating that these represent three independent examples of frequent chromosomal integration by an ESBL.

Consistent with this observation of NCBI761, we were unable to transfer carbenicillin resistance by transformation in 83 non-conjugative isolates from USWest352 containing CG1 or CG2 genes (Supplementary Material). This strongly suggests a chromosomal location of β-lactamase genes in these isolates (not shown).

These observations are also consistent with our linear regression analysis, with resistance to β-lactams exhibiting higher correspondence to USWest352 population structure (mean AUC: 0.79) than the aminoglycosides (mean AUC: 0.68), and higher correspondence than would presumably be expected for genes primarily encoded in plasmids.

Based on these observations, we extend to our model for MDR evolution to include a component for chromosomal integration. While plasmids appear to be the primary platform for generating adaptive solutions, the large observed gene flow to the chromosome, including the individual transfer of most resistance genes and co-transfer of gene combinations comprising adaptive solutions, highlights the involvement of the chromosome in the evolution of acquired MDR. Potentially, chromosomal copies of resistance genes serve as reservoirs for transfer into different plasmids, which would increase the capacity to generate different adaptive solutions.

Further, the large disparity in the degree of chromosomal integration for individual resistance genes (0.0% to 42.5%) likely reflects differential selective pressures between them, with stronger and consistent selective pressure (presumably for the ESBLs in our datasets) favoring chromosomal integration. The selective advantage may be linked to reducing the probability of gene loss because physical linkages to other necessary genes on the bacterial chromosome provide more stable transmission.

Alternatively, this disparity may also reflect differential gene mobility (as in the degree of mobility conferred by specific MGEs associated with individual genes [15]). However, it seems unlikely that highly recombinant MGEs alone could account for the frequent integration of three unlinked ESBL genes, since this would require the coincidental captures of three unlinked ESBL gene by the most highly recombinant MGEs, against the odds that these MGEs stochastically captured other resistance genes. A more parsimonious explanation is that selective pressures drive chromosomal integration.

1.2.16 Concluding remarks

We have identified two groups of genes, CG1 and CG2, that largely explain the distribution of ESBL, gentamicin, and tobramycin resistance, both in the entire set of completed *E. coli* genomes in the NCBI database and in a single epidemic strain: ST131. CG1 and CG2 appear to represent two different adaptive solutions to similar drug selections. The strong β-lactamase representation in both CGs and

the frequent integration of β-lactamases into the chromosome is likely driven by the frequent use of these antibiotics in the clinic for treatment of *E. coli* infection [6,33, 34]. Selection for aminoglycoside resistance (the other class of antibiotic resistance genes represented in CG groups) is likely driven by combination therapy of β-lactam and aminoglycosides, which is prescribed for severe infections because the mechanism of action of these two groups of antibiotics is different [68–70].

CGs can be included in MRRs representing more complex adaptive solutions. They are surprisingly heterogeneous in their specific composition across MLSTs and even within a given MLST. The linkage between the individual genes in each group is highly variable as well. These two observations highlight the key role of selection (as opposed to genetic linkage) in maintaining a diversity of arrangements for each winning solution. We propose that over time, these arrangements tend to converge on solutions that are highly diversified, with genes in close proximity to each other, and sharing the same orientation.

We also found mutual antagonism across resistance genes with overlapping substrates both within the three complementarity groups that we identified, across them, and more generally across the NCBI761, constraining evolution. In combination with co-selection due to physical linkage, this mutual antagonism leads to strong contingency effect that restricts evolutionary trajectories.

We confirm the central role of IncF plasmids in maintaining and spreading antibiotic resistance and as a genetic platform for evolving MDR, but also find that resistance genes, particularly *blaCTX-M* and *blaCMY-2* β-lactamases, frequently integrate into the chromosome. This suggests that vertical transmission through the chromosome may facilitate the spread of resistance genes under consistent selective pressure and that chromosomally integrated drug resistance genes may also serve as a reservoir for transfer into plasmids.

We also present a linear regression model that predicts ESBL, gentamicin and tobramycin resistance with considerable accuracy (AUCs between 0.92 and 0.96). A high concordance between phenotypic and genome-based predictions of antimicrobial susceptibilities has been reported before in the context of resistance surveillance, although in these studies, population structure is not usually factored in [32]. More recently, clinical studies are showing promising results [29]. Our results support the feasibility of routine genotypic prediction of bacterial antimicrobial susceptibility, an approach that looks extremely promising but that is limited by the complexity of mechanisms of antibiotic resistance [71] [72].

Specifically, two aspects of this work have important translational implica-

tions: (1) they show that in the two hospitals that we studied population structure has only a moderate impact on linear regression because of the high level of gene flow for acquired resistance genes; this means that models that ignore population structure may be substantially accurate, at least for surveillance purposes; (2) they identify combinations of genes that dominate the resistance landscape, explaining large portions of the existing genetic variation; further, these combinations appear not to be restricted to a particular location. The observation reported here that the evolution of complex combinations of acquired resistance genes is strongly restricted is of direct relevance for point-of-care diagnostics. However, any genomics-based predictive application to the clinic would need to meet very stringent sensitivity and specificity standards.

1.3 Methods

1.3.1 Sample and data collection

Clinical samples were collected from patients with respiratory, blood (wound), or urinary tract infections at Dignity Health Mercy Medical Center (DHMMC) in Merced, California, between June 2013 and August 2015. The isolates were tested for ESBL resistance using an automated rapid detection system for pathogen identification and antibiotic sensitivity, Vitek 2 Version 06.01. Following identification, the samples were also tested for susceptibility against 16 antibiotics using broth micro-dilution minimum inhibitory concentration (MIC) testing. The isolates were categorized according to their susceptibility: Resistant (R), Intermediate (I), or Susceptible (S), based on the MIC Interpretation Guideline – CLSI M100-S26 (2015). The 16 antibiotics included 1 penicillin: Ampicillin, 2 penicillin and inhibitor combinations: Ampicillin/Sulbactam, Piperacillin/Tazobactam, 4 cephalosporins: Cefazolin, Ceftazidime, Ceftriaxone, Cefepime, 2 carbapenems: Ertapenem, Imipenem, 3 aminoglycosides: Amikacin, Gentamicin, Tobramycin, 2 FQNs: Ciprofloxacin, Levofloxacin, and Nitrofurantoin and Trimetroprim/Sulfamethoxazole. For these antibiotics, the ratios of resistance to susceptible samples that we collected are shown in **Figure 1.10**. Additionally, we obtained 384 ExPEC genome assemblies from a previous study conducted at the University of Washington (UW), downloaded from Genbank with accessions in the range JSFQ00000000–JSST00000000.

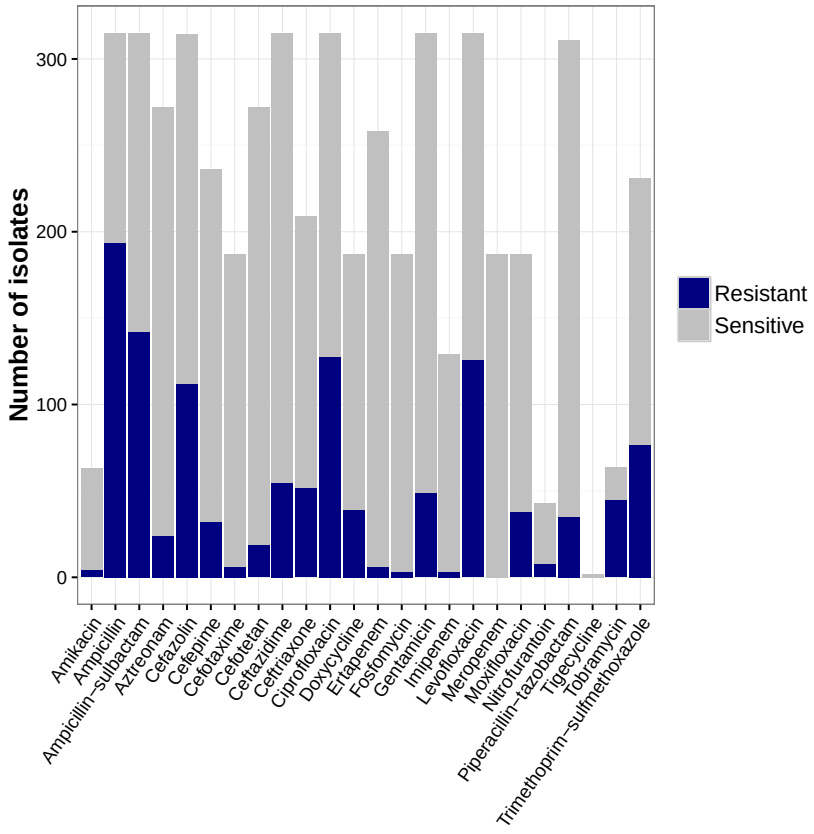

Figure 1.10: Antibiotic susceptibility of the USWest352 isolates. The number of isolates with antibiotic susceptibility testing data is indicated as gray bars, and the proportions of resistant isolates are indicated in blue.

1.3.2 Sequencing, quality control and assembly of ExPEC genomes

Genomic DNA was extracted from each DHMMC sample using the ZR-96 Quick-gDNA Kit from Zymo Research. Whole-genome sequencing including TruSeq DNA library preparation was performed at the University of California, Davis Genome Center using Illumina's MiSeq and HiSeq technologies, and at the University of California, Berkeley using HiSeq. We obtained 24 MiSeq (2x250 bp) and 110 HiSeq (2x250 bp) paired-end sequencing libraries corresponding to our selected samples. Prior to assembly, Illumina sequencing adapters and low quality bases were trimmed from the sequencing reads using Trimmomatic v0.36 [**?**]; trimmed reads shorter than 36 bp were discarded. Library quality was also verified using FastQC v0.11.5. De novo paired-end assembly was conducted for each MiSeq and HiSeq library using SPAdes v3.5.0 [73] with read error correction by BWA-spades. The spades.py wrapper script was used to select an appropriate k-mer size for optimized assembly of each genome. The assemblies had 29x and 327x median

coverage for MiSeq and HiSeq libraries, respectively. The median N50 across 22 MiSeq and 110 HiSeq library assemblies was 198 kb (median L50: 9) and 201 kb (median L50: 8.5).

The resulting assemblies were scanned for contamination. Estimated genome sizes for our libraries (median: 5.3 Mb) were generally within the range of known ExPEC genomes. Several libraries appeared to contain multiple different organisms based on estimated genome sizes (for instance, samples 328 and 357 were 8.2 Mb and 9.0 Mb, respectively), and based on the presence of 16S rDNA genes corresponding to multiple different species; these assemblies were removed from downstream analyses. We further assessed each DHMMC and UW assemblies for completeness based on the presence of 143 protein-coding genes considered to be essential for normal growth of *E. coli* MG1655 (Set A in Supplementary Table 2 from [74]). To achieve consistent assembly qualities across the DHMMC and UW datasets, we removed assemblies containing fewer than 126 full-length essential genes.

1.3.3 Phylogenetic classification of strains

Phylogenetic classification of the 352 draft genome assemblies of USWest352 and 761 completed genome assemblies of NCBI761 was performed using three methods that provided varying levels of resolution: the EzClermont phylotyping method, the Achtman multi-locus sequence typing (MLST) method [72], and based on nucleotide-level variation across whole-genomes. To estimate phylogenetic relationships among the USWest352 genomes, an all-by-all pairwise distance matrix was constructed, in which the genetic distance between each pairing of genomes was estimated as the number of nucleotide positions varying between the two genomes. Variant calling for each genome was performed using the *nesoni consensus* script from the nesoni package. The distance matrix was generated using the *nesoni nway* script from the nesoni package, specifying *E. coli* EC958 as the reference genome [75]. A neighbor-joining tree was constructed based on the distance matrix using SplitsTree4 [76].

1.3.4 Gene model prediction and functional annotation

Gene model predictions were generated using Prodigal v2.6.2 [77], running in metagenomic mode. Genes known to confer resistance against aminoglycosides, β-lactams and FQNs were identified based on amino acid sequence homology (99%

query coverage and greater than 90% sequence identity) to gene entries in the ResFinder database, using NCBI BLASTx and BLASTp [78]. The ResFinder database provided high sequence diversity for a given resistance gene (i.e. sequence variants), so that most hits were distinguishable at 100% query coverage and 100% sequence identity. Full-length hits with less than 100% sequence identity were labeled with nomenclature indicating specific amino acid substitutions e.g., G238S. Hits with less than 99% query coverage, but higher than 90% sequence identity, were closely inspected for completeness. That is, short truncations (10 amino acids or less) at the N-terminus were allowed if this resulted in a new start codon (Methionine or Valine), while hits that appeared to contain larger deletions were considered to be incomplete sequencing/assemblies or true truncations leading to loss of function.

Plasmid replication origins and associated incompatibility groups were annotated based on BLASTn hits (e-value: 10^{-50}) against INC-DB [79], a plasmid origin nucleotide sequence database. Genes encoding plasmid replication initiation proteins, or replicases, and conjugative relaxases were detected using BLASTx and BLASTp against the RIP-DB and REL-DB, respectively, using an e-value threshold of 10^{-50}.

1.3.5 Determination of resistance marker predictive values using logistic regression models, and controls for the confounding effects of population structure

The predictive values of resistance markers identified by our annotation pipeline were determined using logistic regression models, similar to methods employed by recent microbial genome-wide association studies (GWAS).

Various confounding effects can arise from bacterial population structures in association studies [35]. For one, the haploid inheritance of the bacterial chromosome causes linkage disequilibrium on much larger regions than in humans, and long-range linkage disequilibrium in bacterial genealogies can persist despite frequent homologous recombination events. Also, clonal expansions can be driven by positive selection of the phenotype of interest, resulting in a biased distribution of genetic elements that cause the phenotype. These confounding effects can cause false positives for the identification of novel causal variants, or in our case, for assessing the predictive value of causal genetic elements.

Recent bacterial GWAS have sought to develop new methods to counter these confounding effects of population structure with some success. One approach

involved capturing the genealogy represented in a dataset using principal component analysis or multidimensional scaling, so that a subset of the resulting principal components can be included as covariates for regression [36, 80]. Another study showed that modeling population-level effects directly using linear mixed models can greatly increase statistical power for identification of antibiotic resistance markers [81].

Here, we employed the first approach: using multidimensional scaling to capture the genealogical information present in our dataset of 352 *E. coli* genomes. We selected a subset of principal components that captured the major structural patterns in our dataset, and included them as the only regressors in our logistic regression model to establish a baseline for the contribution of population structure to resistance. Subsequently, we identified the contribution of resistance markers to resistance beyond this baseline threshold. We performed classical multidimensional scaling (cmdscale package in R) on the all-by-all ($n \times n$, $n = 352$) nucleotide distance matrix generated using the nesoni package as described in section **1.3.3**.

It has been shown that the number of principal components required to adequately capture the major structural patterns in a given dataset can vary depending on the size of the dataset, and the represented diversity [36]. Including fewer principal components increases sensitivity at the expense of specificity, while including a larger number of components incurs loss of sensitivity, and risks the causal genetic elements being represented in the regressors intended for population structure control.

Two types of plots were used to estimate the optimal number of principal components to retain k, that is specific to our dataset: (1) a stress plot which measures the strain generated by regressing the sample data points after dimensionality reduction onto the original distance matrix (stress is measured as $1–R^2$), and (2) a scree plot that maps the explained variance in the reduced dimensional space (i.e. eigenvalues) against the number of dimensions (**Figure 1.8**). Based on a visual inspection of these two plots, the first three principal components ($k = 3$) appeared to represent a good trade-off between sensitivity and specificity. Our choice of $k = 3$ was supported by Catell's scree test [82], and by a previous study that performed GWAS on a dataset of similar size [36]. Thus we retained a set of control regressors $P = \{P_1, P_2, ..., P_k\}$ for $k = 3$.

The antibiotic resistance markers identified by our annotation pipeline represented a second set of regressors, M, in our model. These included a total of $l = 38$ regressors corresponding to 34 acquired aminoglycoside, β-lactam and fluoroquinolone resistance genes occurring in at least one of the 352 genomes, and also 4

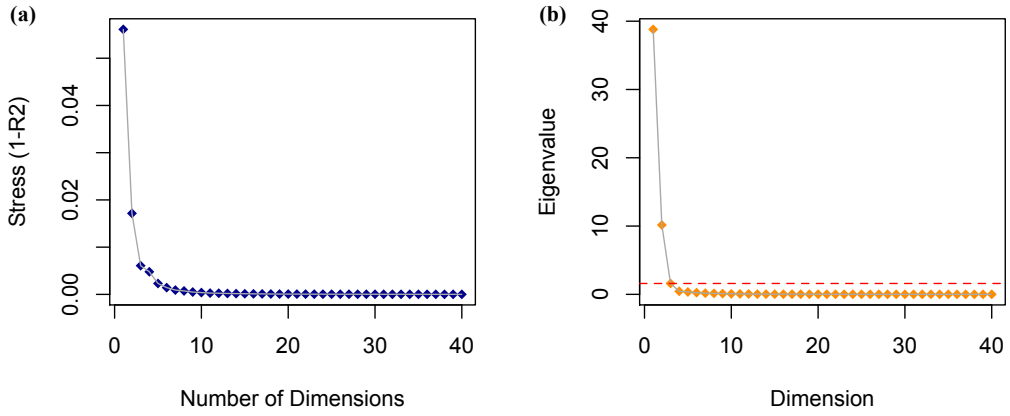

Figure 1.11: Stress and scree plots. The two plots were used to estimate an ideal number of principal components to retain as population structure control regressors. **(a)** Stress plot shows the strain generated from regressing the distance between sample data points after dimensionality reduction to the original set of distances. **(b)** Scree plot maps the explained variance against the number of dimensions. The dotted redline indicates the number of statistically significant factors ($k = 3$), as determined by Catell's test.

mutations in chromosomal Topoisomerase IV genes known to confer fluoroquinolone resistance. The presence or absence of marker $j \in \{$*aac(3)-IIa*, *aac(3)-IId*, *aac(3)-VIa*, *aac(6')-Ib-cr*, *aadA1*, *aadA16*, *aadA1b*, *aadA2b*, *aadA5*, *ant(2")-Ia*, *aph(3")-Ib*, *aph(3')-Ia*, *aph(6)-Id*, *rmtE*, *blaCARB-11*, *blaCARB-2*, *blaCMY-2*, *blaCTX-M-1*, *blaCTX-M-104*, *blaCTX-M-14b*, *blaCTX-M-15*, *blaCTX-M-27*, *blaCTX-M-55*, *blaCTX-M-65*, *blaOXA-1*, *blaTEM-12*, *blaTEM-19*, *blaTEM-1A*, *qepA4*, *qnrA1*, *qnrB19*, *qnrB6*, *qnrS2*, *gyrA*-S83X, *gyrA*-D87X, *parC*-S80X, *parC*-E84X$\}$ in the ith genome is denoted as $m_{i,j} \in \{0, 1\}$ for $i \in \{1, ..., n\}$ and $j \in \{1, ..., l\}$.

Finally, the R, I and S resistance phenotypes were modeled as a binary response variable $y_{i,d} \in \{0, 1\}$ for logistic regression, with both R (high-level resistance) and I (intermediate resistance) being assigned a value of 1, for the set of drugs $d \in \{$AMP, CAZ, CIP, CRO, CFZ, FEP, GEN, LVX, TOB$\}$.

Logistic regression was performed in three different modes: (mode 1) for establishing the baseline contribution of population structure to resistance ($X = [P_{(n\times k)}]$), (mode 2) for estimating the increase in predictive accuracy conferred by the set of resistance markers ($X = [P_{(n\times k)}\ M_{(n\times l)}]$), and (mode 3) for confirming that the resistance markers alone were sufficient to achieve high prediction accuracy without the population control regressors ($X = [M_{(n\times l)}]$). The predictive accuracy of a given markers j was estimated based on the magnitude of decrease in predictive accuracy conferred by the removal of marker j from the model in mode 2.

Penalized logistic regression was performed as implemented in the glmnet package for R [83]. The objective function is provided here as reference (**Eq. 1.1**).

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\sum_{i=1}^{n} y_{i,d}\cdot(\beta_0 + x_{i,d}^T\beta) - \log\left(1+e^{(\beta_0+x_{i,d}^T\beta)}\right)\right] + \lambda\left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right] \quad (1.1)$$

In **Eq. 1.1**, β and β_0 are the learned coefficients and intercepts, respectively, and x_i denotes the binary presence or absence of markers in sample genome i. The glmnet package estimates a quadratic approximation to the negative log-likelihood, and performs gradient descent to solve the resulting least squares problem. A wrapper script was written to execute the cv.glmnet function using five-fold cross validation, and with elastic net regularization ($\alpha = 0.5$), to prevent degenerate behavior characteristic of logistic regression and to deal with possible multicollinearity. In glmnet, the amount of regularization used is controlled by the λ parameter, which is learned at each execution of cv.glmnet. As a heuristic, λ returning the minimum mean cross-validated error after cross validation was chosen at each execution of cv.glmnet. To further discourage overfitting, cv.glmnet was executed on 5000 bootstrap replicates, sampling 80% of the 352 genomes at each bootstrap iteration, while the remaining 20% was used for validation.

1.3.6 Exploratory factor analysis using a bootstrapped MDS procedure

To identify global patterns in the distribution of acquired resistance genes identified by our annotation pipeline in USWest352, we performed exploratory factor analysis using multidimensional scaling (MDS).

MDS is a popular method for information visualization with wide applications. However, in standard applications, MDS outcomes are provided without statistical support, and thus can be misleading. Some studies have proposed statistical interpretations of MDS outcomes based on bootstrap sampling, by generating confidence intervals for the projected data coordinates [84].

Here, we implemented MDS using a bootstrapping procedure to improve the accuracy and reproducibility of the final MDS outcome. Bootstrapping was performed by randomly selecting 80% of the genomes without replacement. For each bootstrap replicate, we performed unsupervised classification of the MDS outcome, based on the distribution of projected data coordinates in Euclidean space. A final MDS outcome was generated using a distance matrix P that provided the pair-wise probabilities of two given genomes being placed in different clusters. For genomes i

and j ($i \neq j$), the probability of being placed in different clusters was calculated as $p_{ij} = 1 - g_{ij}/h_{ij}$, where g_{ij} represents the number of times that i and j were placed in the same cluster, and h_{ij} represents the number of times that both i and j were selected in the same bootstrap replicate.

Classical MDS was implemented using the cmdscale package for R, which requires a distance matrix as input. The Jaccard distance (**Eq. 1.2**) was used to model the dissimilarity in resistance gene composition between two given genomes, based on the presence or absence of 34 aminoglycoside, β-lactam and fluoroquinolone resistance genes identified by our annotation pipeline.

$$\text{Jaccard Dist.} = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{1.2}$$

In **Eq. 1.2**, A represents the set of resistance genes present in genome i, and B represents the set of resistance genes present in genome j.

Unsupervised classification was performed using the mclust package for R [45], which uses gaussian mixtures to model data clusters with parameter estimation by expectation maximization. For each bootstrap replicate, the optimal number of clusters k was approximated using the Bayesian Information Criterion (BIC). The k with the highest BIC was chosen from the range $k = 1$ to $k = 20$.

1.3.7 Significance testing for gene co-occurrence using a random distribution model

We identified gene co-occurrences that are significantly overrepresented or underrepresented in USWest352, CLONAL161 and NCBI761, by performing significance testing against a random distribution model.

The model was constructed as follows. Given a dataset S containing n genomes, the collection of ORF hits corresponding to resistance genes found in S can be represented as a gene pool G, so that the composition of P reflects the empirical frequencies of individual genes represented in S. Additionally, each genome $i \in \{1, \ldots, n\}$ can be assigned a genome size t_i equivalent to the number of ORF hits it contributes to G. To obtain a permuted genome p_i, we randomly sampled t_i genes from G without replacement. A full set of n permuted genomes corresponding to dataset S is represented by $P \in \{p_1, p_2, \ldots, p_n\}$.

We assessed the significance of co-occurrence for gene pairs using an empirical p-value calculated as shown in (**Eq. 1.3**), against the two-tailed null hypotheses shown in (**Eq. 1.4**).

$$\text{p-value} = 1 - \frac{\text{No. of times } H_0 \text{ is rejected}}{\text{No. of permuted datasets}} \tag{1.3}$$

$$H_0 : c_{u,v}(S) = c_{u,v}(P) \tag{1.4}$$

In **Eq. 1.4**, $c_{u,v}(S)$ denotes the observed co-occurrence of genes u and v in dataset S, and $c_{u,v}(P)$ denotes their co-occurrence in a permuted dataset P. The co-occurrence of genes u and v was measured using the Jaccard index (**Eq. 1.5**).

$$c_{u,v}(X) = 1 - \frac{|X_u \cap X_v|}{|X_u \cup X_v|} \tag{1.5}$$

In **Eq. 1.5**, X_u is the set of genomes from dataset X that contain gene u, and X_v is the set of genomes from dataset X that contain gene v.

P-values for overrepresentation and underrepresentation of co-occurrence were generated using 5000 permuted datasets, and were adjusted for multiple comparisons using the false discovery rate at a significance level of $\alpha = 0.05$.

1.3.8 Modified significance test to detect negative co-selection within gene groups

We performed a modified version of the significance test to detect negative co-selection within groups of genes with similar functions. The 108 genes found in NCBI761 were assigned to $k = 13$ functional groups. The number of ORF hits corresponding to each functional group (sample size) is provided in **Table 1.5**. The gene-to-group assignments for specific genes are provided at the end of this section. The modified null hypothesis is provided below.

$$H_0 : m_j(S) \geq m_j(P) \tag{1.6}$$

In **Eq. 1.6**, $m_j(S)$ and $m_j(P)$ denote the number of genomes in datasets S and P, respectively, that contain two or more genes from functional group $j \in \{1, \ldots, k\}$. P-values were generated for the 13 functional groups against the modified null hypothesis using 1000 permuted datasets, and assessed at a significance level of $\alpha = 0.05$.

We conducted an additional test to verify that we had captured meaningful gene-to-group assignments, and that our set of p-values corresponding to the 13 functional groups was not obtained by chance of random assignment. The modified significance test was performed an additional 1000 rounds, generating 1000 sets of p-values, and during which we randomly permuted the gene-to-group assignments for the 13 functional groups. Each round was supported by 1000 datasets containing permuted gene co-occurrences. We calculated a grand-p-value to assess the

significance of our original set of p-values obtained with the "true" gene-to-group assignments (**Eq. 1.7**), against the null hypothesis shown in (**Eq. 1.8**).

$$\text{p-value} = 1 - \frac{\text{No. of times } H_0 \text{ is rejected for} k \text{groups}}{\text{No. of gene-to-group assignment permutation rounds}} \tag{1.7}$$

$$H_0 : \text{p-value}_j^{true} \geq \text{p-value}_j^{perm} \tag{1.8}$$

In **Eq. 1.8**, p-value_j^{true} and p-value_j^{perm} represent the p-values obtained from the "true" gene-to-group assignments and permuted assignments, respectively, for functional group $j \in \{1, \ldots, k\}$.

1.3.9 NCBI761 functional gene group assignments

The gene-to-group assignments for 108 different aminoglycoside, β-lactam and fluoroquinolone resistance genes identified in NCBI761 are provided below. These 108 genes were collected from 1802 distinct genetic loci, and corresponded to full-length ORF hits.

Genes encoding AMEs were divided among five functional groups based on their known resistance spectrums, while the 16S rRNA methyltransferase genes were placed in a separate group.

Group 1 included *aac(3)-IIa*, *aac(3)-IId*, *aac(3)-IV*, *aac(3)-Ia*, *aac(3)-VIa* and *ant(2")-Ia*, most of which encode 3-N-acetyltransferases that confer resistance against GEN and TOB, but not AMK [85]. Several of the members (*aac(3)-IIa*, *aac(3)-IId* and *aac(3)-VIa*) appear to have lower activity against TOB than to GEN [86], while *aac(3)-Ia* does not confer TOB resistance [85]. Notably, *aac(3)-IV* has a broad resistance spectrum against aminoglycosides including GEN, TOB, AMK and apramycin [50], and did not fit well into Group 1 or Group 2. We included *aac(3)-IV* in Group1 based on its high resistance against GEN, and based no its method of catalysis i.e. it encodes an 3-N-acetyltransferase. On the other hand, *ant(2")-Ia* does not encode an N-3-acetyltransferases, but was assigned to Group 1 because it confers resistance against GEN and TOB, but not AMK [87].

Group 2 included *aac(6')-33*, *aac(6')-Ian*, *aac(6')-Ib*, *aac(6')-Ib-cr*, *aac(6')-Ib3* and *aac(6')-Il*, which encode type I 6'-N-acetyltrasnferases that confer resistance against TOB and AMK, but not GEN [57].

Group 3 included *aadA1*, *aadA13*, *aadA16*, *aadA2*, *aadA22*, *aadA23*, *aadA24*, *aadA2b*, *aadA3*, *aadA5*, which encode ANT(3")-Ia variants that confer resistance against STR and SPC, but not GEN, TOB or AMK [20, 57].

Group 4 included *aph(3")-Ib* and *aph(6)-Id*, which are also called strA and strB, respectively, confer STR resistance only. These two genes have been found closely linked in a wide range of species [10, 85], and within a STR resistance operon in Shigella flexneri [88].

Group 5 included *aph(3')-IIa*, *aph(3')-Ia*, *aph(3')-VI*, *aph(3')-VIa* and *aph(3')-VIb*, which encode 3'-O-phosphotransferases that confer resistance against kanamycin, neomycin and pradimicin but not GEN, TOB, STR or SPC [85].

Group 6 included three genes encoding 16S rRNA methyltransferases, armA, rmtB and rmtC.

Genes encoding β-lactamases were divided into five groups, approximately based on function. We found that classification of the β-lactamases by subclass/family (e.g., the CMYs, CTX-Ms, KPCs, NDMs, OXAs and TEMs) provided adequate distinction based on resistance spectrum, focusing on resistance against the cephamycins, $3^{\text{rd}}/4^{\text{th}}$-generation cephalosporins, aztreonam and the carbapenems. Some exceptions, if known, are stated in the following.

Group 7 included *blaCMY-2*, *blaCMY-4*, *blaCMY-6*, *blaCMY-16*, *blaCMY-24*, *blaCMY-34*, *blaCMY-42*, *blaCMY-44*, *blaCMY-111* and *blaCMY-new*, which are plasmid-borne AmpC β-lactamases that confer resistance against 3^{rd}-generation cephalosporins, some cephamycins, and meropenem. It has been reported that *blaCMY-2* has the capability of acquiring resistance against the 4^{th}-generation cephalosporin cefepime through mutation [89].

Group 8 included *blaCTX-M-2*, *blaCTX-M-3*, *blaCTX-M-14b*, *blaCTX-M-15*, *blaCTX-M-24*, *blaCTX-M-27*, *blaCTX-M-55*, *blaCTX-M-64*, *blaCTX-M-65*, *blaCTX-M-123*, *blaCTX-M-199*, which generally confer ESBL resistance and resistance against aztreonam, but do not confer resistance against carbapenems or cephamycins. Notably, *blaCTX-M-14b* is reported to have poor activity against CAZ and aztreonam [90].

Group 9 included *blaKPC-2*, *blaKPC-3*, *blaKPC-4*, which mainly confers carbapenem resistance, and resistance against penicillins and first generation cephalosporins. Members of this group can confer resistance against some cephamycins e.g., cefoxitin for *blaKPC-3* [28, 91, 92].

Group 10 included *blaNDM-1*, *blaNDM-4*, *blaNDM-5*, *blaNDM-6*, *blaNDM-7*, *blaNDM-9*, and *blaNDM-21*, which are carbapenemases that can confer resistance against virtually all β-lactams except for aztreonam. Additionally, *blaNDM-1* has low catalytic efficiency against CAZ [7].

Group 11 included *blaOXA-1*, *blaOXA-2*, *blaOXA-4*, *blaOXA-9*, *blaOXA-*

10, *blaOXA-48*, *blaOXA-163*, *blaOXA-181*, which all confer resistance against penicillins and oxacillins. However, *blaOXA-48*, *blaOXA-163* and *blaOXA-181* additionally confer resistance against carbapenems [93], and comprise a significant minority. Thus this group is functionally heterogeneous in this regard. Other OXA variants that confer ESBL resistance has been found in *P. areuginosa*, but we did not find any of these in NCBI761 [93].

Group 12 included *blaTEM-1A*, *blaTEM-20*, *blaTEM-26*, *blaTEM-30*, *blaTEM-32*, *blaTEM-57*, *blaTEM-116*, *blaTEM-135*, *blaTEM-156*, *blaTEM-176*, *blaTEM-210*, *blaTEM-215*. Most members of this group confer resistance against penicillins and narrow spectrum cephalosporins except *blaTEM-1A*, which only confers resistance against penicillins. The variants *blaTEM-20* and *blaTEM-26* also confer ESBL resistance, but are found in only two genomes in NCBI761.

Group 13 was comprised of 12 PMQR genes (*qpepA1*, *qpepA4*, *qnrA1*, *qnrB4*, *qnrB6*, *qnrB9*, *qnrB10*, *qnrE1*, *qnrS1*, *qnrS2* and *qnrVC4*), which confer plasmid-mediated resistance against fluoroquinolones.

Some genes encoding AMEs and β-lactamases that were placed in additional groups were removed from this analysis due to low sample size. This included the AME genes *aac(2')-IIa* and *aph(4)-Ia*, which confer resistance against kasugamycin and hygramycin, respectively, and the IMP, VEB and VIM carbapenemases.

1.3.10 Detection of the transferability of resistance genes by conjugation assays

Horizontal transferability of resistance genes belonging to CG1 and CG2 was detected by conjugation assays using *E. coli* LMB100 (donated by Dr. Luis Mota-Bravo) as the recipient strain. The assay was performed on 146 ExPEC isolates (39 from UW233 and 107 from DHMMC119).

Transconjugants were selected on MacConkey agar plates supplemented with carbenicillin (100 μg/ml) and rifampicin (100 μg/ml) and subsequently characterized. Transferred plasmids were classified according to their incompatibility group using the classic PCR-based replicon typing method (Carattoli et al. 2005) (Table A). Plasmids from donors and transconjugants were visualized by pulsed-field gel electrophoresis (PFGE) with S1 nuclease (Thermo Fisher Scientific) digestion, and plasmid sizes were estimated by comparing the MidRange I PFG Marker (New England Biolabs) through the least squares method (Statgraphics 18 software). The receptor *E. coli* LMB100 was used as a negative control in PFGE experiments

(plasmid-free strain).

1.3.11 Determination of phenotypic specificity of AMEs

To investigate the phenotypic specificity of three AMEs (*aac(3)-IIa*, *aac(3)-IIa* and *aac(6')-Ib-cr*), we tested isolates harboring these genes for susceptibility to TOB and GEN using the disk diffusion method (disks containing 10 μg of antibiotic). Results were interpreted as specified by the Clinical and Laboratory Standards Institute guidelines.

Bibliography

[1] S. Gandra, K. K. Tseng, A. Arora, B. Bhowmik, M. L. Robinson, B. Panigrahi, R. Laxminarayan, and E. Y. Klein. The mortality burden of multidrug-resistant pathogens in india: a retrospective observational study. *Clin Infect Dis*, 2018.

[2] R. E. Nelson, R. B. Slayton, V. W. Stevens, M. M. Jones, K. Khader, M. A. Rubin, J. A. Jernigan, and M. H. Samore. Attributable mortality of healthcare-associated infections due to multidrug-resistant gram-negative bacteria and methicillin-resistant staphylococcus aureus. *Infect Control Hosp Epidemiol*, 38(7):848–856, 2017.

[3] E. Sauvage and M. Terrak. Glycosyltransferases and transpeptidases/penicillin-binding proteins: Valuable targets for new antibacterials. *Antibiotics (Basel)*, 5(1), 2016.

[4] K. Z. Vardakas, P. I. Rafailidis, A. A. Konstantelias, and M. E. Falagas. Predictors of mortality in patients with infections due to multi-drug resistant gram negative bacteria: the study, the patient, the bug or the drug? *J Infect*, 66(5):401–14, 2013.

[5] M. Exner, S. Bhattacharya, B. Christiansen, J. Gebel, P. Goroncy-Bermes, P. Hartemann, P. Heeg, C. Ilschner, A. Kramer, E. Larson, W. Merkens, M. Mielke, P. Oltmanns, B. Ross, M. Rotter, R. M. Schmithausen, H. G. Sonntag, and M. Trautmann. Antibiotic resistance: What is so special about multidrug-resistant gram-negative bacteria? *GMS Hyg Infect Control*, 12:Doc05, 2017.

[6] D. Koulenti, A. Song, A. Ellingboe, M. H. Abdul-Aziz, P. Harris, E. Gavey, and J. Lipman. Infections by multidrug-resistant gram-negative bacteria: What's new in our arsenal and what's in the pipeline? *Int J Antimicrob Agents*, 53(3):211–224, 2019.

[7] D. Yong, M. A. Toleman, C. G. Giske, H. S. Cho, K. Sundman, K. Lee, and T. R. Walsh. Characterization of a new metallo-beta-lactamase gene, bla(ndm-1), and a novel erythromycin esterase gene carried on a unique genetic structure in klebsiella pneumoniae sequence type 14 from india. *Antimicrob Agents Chemother*, 53(12):5046–54, 2009.

[8] S. J. Salipante, D. J. Roach, J. O. Kitzman, M. W. Snyder, B. Stackhouse, S. M. Butler-Wu, C. Lee, B. T. Cookson, and J. Shendure. Large-scale genomic sequencing of extraintestinal pathogenic escherichia coli strains. *Genome Res*, 25(1):119–28, 2015.

[9] H. Nikaido. Multidrug resistance in bacteria. *Annu Rev Biochem*, 78:119–46, 2009.

[10] M. Ashenafi, T. Ammosova, S. Nekhai, and W. M. Byrnes. Purification and characterization of aminoglycoside phosphotransferase aph(6)-id, a streptomycin-inactivating enzyme. *Mol Cell Biochem*, 387(1-2):207–16, 2014.

[11] R. Canton. Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. *Clin Microbiol Infect*, 15 Suppl 1:20–5, 2009.

[12] T. S. Crofts, A. J. Gasparrini, and G. Dantas. Next-generation approaches to understand and combat the antibiotic resistome. *Nat Rev Microbiol*, 15(7):422–434, 2017.

[13] M. Fondi, A. Karkman, M. V. Tamminen, E. Bosi, M. Virta, R. Fani, E. Alm, and J. O. McInerney. "every gene is everywhere but the environment selects": Global geolocalization of gene sharing in environmental samples through network analysis. *Genome Biol Evol*, 8(5):1388–400, 2016.

[14] G. Guedon, V. Libante, C. Coluzzi, S. Payot, and N. Leblond-Bourget. The obscure world of integrative and mobilizable elements, highly widespread elements that pirate bacterial conjugative systems. *Genes (Basel)*, 8(11), 2017.

[15] S. R. Partridge. Analysis of antibiotic resistance regions in gram-negative bacteria. *FEMS Microbiol Rev*, 35(5):820–55, 2011.

[16] M. N. Alekshun and S. B. Levy. Molecular mechanisms of antibacterial multidrug resistance. *Cell*, 128(6):1037–50, 2007.

[17] M. Vos, M. C. Hesselman, T. A. Te Beek, M. W. J. van Passel, and A. Eyre-Walker. Rates of lateral gene transfer in prokaryotes: High but why? *Trends Microbiol*, 23(10):598–605, 2015.

[18] F. Hu, J. A. O'Hara, J. I. Rivera, and Y. Doi. Molecular features of community-associated extended-spectrum-beta-lactamase-producing escherichia coli strains in the united states. *Antimicrob Agents Chemother*, 58(11):6953–7, 2014.

[19] J. Shin, M. J. Choi, and K. S. Ko. Replicon sequence typing of incf plasmids and the genetic environments of blactx-m-15 indicate multiple acquisitions of blactx-m-15 in escherichia coli and klebsiella pneumoniae isolates from south korea. *J Antimicrob Chemother*, 67(8):1853–7, 2012.

[20] M. S. Ramirez and M. E. Tolmasky. Aminoglycoside modifying enzymes. *Drug Resist Updat*, 13(6):151–71, 2010.

[21] Y. Doi, Y. S. Park, J. I. Rivera, J. M. Adams-Haduch, A. Hingwe, E. M. Sordillo, 2nd Lewis, J. S., W. J. Howard, L. E. Johnson, B. Polsky, J. H. Jorgensen, S. S. Richter, K. A. Shutt, and D. L. Paterson. Community-associated extended-spectrum beta-lactamase-producing escherichia coli infection in the united states. *Clin Infect Dis*, 56(5):641–8, 2013.

[22] F. Baquero. From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nat Rev Microbiol*, 2(6):510–8, 2004.

[23] R. Canton and P. Ruiz-Garbajosa. Co-resistance: an opportunity for the bacteria and resistance genes. *Curr Opin Pharmacol*, 11(5):477–85, 2011.

[24] H. W. Jannasch and M. J. Mottl. Geomicrobiology of deep-sea hydrothermal vents. *Science*, 229(4715):717–25, 1985.

[25] E. Dellus-Gur, M. Elias, E. Caselli, F. Prati, M. L. Salverda, J. A. de Visser, J. S. Fraser, and D. S. Tawfik. Negative epistasis and evolvability in tem-1 beta-lactamase–the thin line between an enzyme's conformational freedom and disorder. *J Mol Biol*, 427(14):2396–409, 2015.

[26] S. R. Partridge, Z. Zong, and J. R. Iredell. Recombination in is26 and tn2 in the evolution of multiresistance regions carrying blactx-m-15 on conjugative incf plasmids from escherichia coli. *Antimicrob Agents Chemother*, 55(11):4971–8, 2011.

[27] D. H. Erwin. Evolutionary contingency. *Curr Biol*, 16(19):R825–6, 2006.

[28] T. Palzkill. Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of tem, ctx-m, and kpc beta-lactamases. *Front Mol Biosci*, 5:16, 2018.

[29] N. Stoesser, E. M. Batty, D. W. Eyre, M. Morgan, D. H. Wyllie, C. Del Ojo Elias, J. R. Johnson, A. S. Walker, T. E. Peto, and D. W. Crook. Predicting antimicrobial susceptibilities for escherichia coli and klebsiella pneumoniae isolates using whole genomic sequence data. *J Antimicrob Chemother*, 68(10):2234–44, 2013.

[30] N. Stoesser, A. E. Sheppard, L. Pankhurst, N. De Maio, C. E. Moore, R. Sebra, P. Turner, L. W. Anson, A. Kasarskis, E. M. Batty, V. Kos, D. J. Wilson, R. Phetsouvanh, D. Wyllie, E. Sokurenko, A. R. Manges, T. J. Johnson, L. B. Price, T. E. Peto, J. R. Johnson, X. Didelot, A. S. Walker, D. W. Crook, and Group Modernizing Medical Microbiology Informatics. Evolutionary history of the global emergence of the escherichia coli epidemic clone st131. *MBio*, 7(2):e02162, 2016.

[31] E. Zankari, H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, and M. V. Larsen. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*, 67(11):2640–4, 2012.

[32] E. Zankari, H. Hasman, R. S. Kaas, A. M. Seyfarth, Y. Agerso, O. Lund, M. V. Larsen, and F. M. Aarestrup. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother*, 68(4):771–7, 2013.

[33] E. Cerceo, S. B. Deitelzweig, B. M. Sherman, and A. N. Amin. Multidrug-resistant gram-negative bacterial infections in the hospital setting: Overview, implications for clinical practice, and emerging treatment options. *Microb Drug Resist*, 22(5):412–31, 2016.

[34] H. M. Zowawi, P. N. Harris, M. J. Roberts, P. A. Tambyah, M. A. Schembri, M. D. Pezzani, D. A. Williamson, and D. L. Paterson. The emerging threat of multidrug-resistant gram-negative bacteria in urology. *Nat Rev Urol*, 12(10):570–84, 2015.

[35] R. A. Power, J. Parkhill, and T. de Oliveira. Microbial genome-wide association studies: lessons from human gwas. *Nat Rev Genet*, 18(1):41–50, 2017.

[36] J. A. Lees, M. Vehkala, N. Valimaki, S. R. Harris, C. Chewapreecha, N. J. Croucher, P. Marttinen, M. R. Davies, A. C. Steer, S. Y. Tong, A. Honkela, J. Parkhill, S. D. Bentley, and J. Corander. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*, 7:12797, 2016.

[37] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 2006.

[38] S. Correia, P. Poeta, M. Hebraud, J. L. Capelo, and G. Igrejas. Mechanisms of quinolone action and resistance: where do we stand? *J Med Microbiol*, 66(5):551–559, 2017.

[39] G. A. Jacoby. Mechanisms of resistance to quinolones. *Clin Infect Dis*, 41 Suppl 2:S120–6, 2005.

[40] N. Tsukamoto, Y. Ohkoshi, T. Okubo, T. Sato, O. Kuwahara, N. Fujii, Y. Tamura, and S. Yokota. High prevalence of cross-resistance to aminoglycosides in fluoroquinolone-resistant escherichia coli clinical isolates. *Chemotherapy*, 59(5):379–84, 2013.

[41] S. Ghafourian, N. Sadeghifard, S. Soheili, and Z. Sekawi. Extended spectrum beta-lactamases: Definition, classification and epidemiology. *Curr Issues Mol Biol*, 17:11–21, 2015.

[42] E. S. Donkor. Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics. *Genes (Basel)*, 4(4):556–72, 2013.

[43] P Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Des. Sci. Nat*, 44:223–270, 1906.

[44] A. J. Mathers, G. Peirano, and J. D. Pitout. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant enterobacteriaceae. *Clin Microbiol Rev*, 28(3):565–91, 2015.

[45] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R J*, 8(1):289–317, 2016.

[46] H. Kanamori, C. M. Parobek, J. J. Juliano, J. R. Johnson, B. D. Johnston, T. J. Johnson, D. J. Weber, W. A. Rutala, and D. J. Anderson. Genomic analysis of multidrug-resistant escherichia coli from north carolina community hospitals: Ongoing circulation of ctx-m-producing st131-h30rx and st131-h30r1 strains. *Antimicrob Agents Chemother*, 61(8), 2017.

[47] S. Mushtaq, S. Irfan, J. B. Sarma, M. Doumith, R. Pike, J. Pitout, D. M. Livermore, and N. Woodford. Phylogenetic diversity of escherichia coli strains producing ndm-type carbapenemases. *J Antimicrob Chemother*, 66(9):2002–5, 2011.

[48] M. H. Wong, E. W. Chan, and S. Chen. Evolution and dissemination of oqxab-like efflux pumps, an emerging quinolone resistance determinant among members of enterobacteriaceae. *Antimicrob Agents Chemother*, 59(6):3290–7, 2015.

[49] P. J. Stogios, T. Shakya, E. Evdokimova, A. Savchenko, and G. D. Wright. Structure and function of aph(4)-ia, a hygromycin b resistance enzyme. *J Biol Chem*, 286(3):1966–75, 2011.

[50] M. L. Magalhaes and J. S. Blanchard. The kinetic mechanism of aac3-iv aminoglycoside acetyltransferase from escherichia coli. *Biochemistry*, 44(49):16275–83, 2005.

[51] M. Bell. Antibiotic misuse: a global crisis. *JAMA Intern Med*, 174(12):1920–1, 2014.

[52] J. E. Mroczkowska and M. Barlow. Recombination and selection can remove blatem alleles from bacterial populations. *Antimicrob Agents Chemother*, 52(9):3408–10, 2008.

[53] S. Paul, S. Million-Weaver, S. Chattopadhyay, E. Sokurenko, and H. Merrikh. Accelerated gene evolution through replication-transcription conflicts. *Nature*, 495(7442):512–5, 2013.

[54] C. M. Miton and N. Tokuriki. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci*, 25(7):1260–72, 2016.

[55] M. L. Salverda, E. Dellus, F. A. Gorter, A. J. Debets, J. van der Oost, R. F. Hoekstra, D. S. Tawfik, and J. A. de Visser. Initial mutations direct alternative pathways of protein evolution. *PLoS Genet*, 7(3):e1001321, 2011.

[56] M. F. Schenk, I. G. Szendro, M. L. Salverda, J. Krug, and J. A. de Visser. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Mol Biol Evol*, 30(8):1779–87, 2013.

[57] P. Shah, D. M. McCandlish, and J. B. Plotkin. Contingency and entrenchment in protein evolution under purifying selection. *Proc Natl Acad Sci U S A*, 112(25):E3226–35, 2015.

[58] T. N. Starr, J. M. Flynn, P. Mishra, D. N. A. Bolon, and J. W. Thornton. Pervasive contingency and entrenchment in a billion years of hsp90 evolution. *Proc Natl Acad Sci U S A*, 115(17):4453–4458, 2018.

[59] L. Ma, Y. Ishii, F. Y. Chang, K. Yamaguchi, M. Ho, and L. K. Siu. Ctx-m-14, a plasmid-mediated ctx-m type extended-spectrum beta-lactamase isolated from escherichia coli. *Antimicrob Agents Chemother*, 46(6):1985–8, 2002.

[60] M. Fondi and R. Fani. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. *Environ Microbiol*, 12(12):3228–42, 2010.

[61] Tom Howley, Michael G. Madden, Marie-Louise O'Connell, and Alan G. Ryder.

[62] S. Shaik, A. Ranjan, S. K. Tiwari, A. Hussain, N. Nandanwar, N. Kumar, S. Jadhav, T. Semmler, R. Baddam, M. A. Islam, M. Alam, L. H. Wieler, H. Watanabe, and N. Ahmed. Comparative genomic analysis of globally dominant st131 clone with other epidemiologically successful extraintestinal pathogenic escherichia coli (expec) lineages. *MBio*, 8(5), 2017.

[63] A. K. Van der Bij, G. Peirano, A. Pitondo-Silva, and J. D. Pitout. The presence of genes encoding for different virulence factors in clonally related escherichia coli that produce ctx-ms. *Diagn Microbiol Infect Dis*, 72(4):297–302, 2012.

[64] G. Koraimann. Spread and persistence of virulence and antibiotic resistance genes: A ride on the f plasmid conjugation module. *EcoSal Plus*, 8(1), 2018.

[65] L. A. Banuelos-Vazquez, G. Torres Tejerizo, and S. Brom. Regulation of conjugative transfer of plasmids and integrative conjugative elements. *Plasmid*, 91:82–89, 2017.

[66] A. Carattoli. Resistance plasmid families in enterobacteriaceae. *Antimicrob Agents Chemother*, 53(6):2227–38, 2009.

[67] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–20, 2014.

[68] G. Cox, P. J. Stogios, A. Savchenko, and G. D. Wright. Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ant(2")-ia. *MBio*, 6(1), 2015.

[69] J. H. Jeon, J. H. Lee, J. J. Lee, K. S. Park, A. M. Karim, C. R. Lee, B. C. Jeong, and S. H. Lee. Structural basis for carbapenem-hydrolyzing mechanisms of carbapenemases conferring antibiotic resistance. *Int J Mol Sci*, 16(5):9654–92, 2015.

[70] M. P. Mingeot-Leclercq, Y. Glupczynski, and P. M. Tulkens. Aminoglycosides: activity and resistance. *Antimicrob Agents Chemother*, 43(4):727–37, 1999.

[71] M. Su, S. W. Satola, and T. D. Read. Genome-based prediction of bacterial antibiotic resistance. *J Clin Microbiol*, 57(3), 2019.

[72] T. J. Johnson, Y. M. Wannemuehler, S. J. Johnson, C. M. Logue, D. G. White, C. Doetkott, and L. K. Nolan. Plasmid replicon typing of commensal and pathogenic escherichia coli isolates. *Appl Environ Microbiol*, 73(6):1976–83, 2007.

[73] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19(5):455–77, 2012.

[74] S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabasi, Z. N. Oltvai, and A. L. Osterman. Experimental determination and system level analysis of essential genes in escherichia coli mg1655. *J Bacteriol*, 185(19):5673–84, 2003.

[75] D. Du, Z. Wang, N. R. James, J. E. Voss, E. Klimont, T. Ohene-Agyei, H. Venter, W. Chiu, and B. F. Luisi. Structure of the acrab-tolc multidrug efflux pump. *Nature*, 509(7501):512–5, 2014.

[76] Y. Doi, J. I. Wachino, and Y. Arakawa. Aminoglycoside resistance: The emergence of acquired 16s ribosomal rna methyltransferases. *Infect Dis Clin North Am*, 30(2):523–537, 2016.

[77] D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, 2010.

[78] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[79] V. F. Lanza, M. de Toro, M. P. Garcillan-Barcia, A. Mora, J. Blanco, T. M. Coque, and F. de la Cruz. Plasmid flux in escherichia coli st131 sublineages, analyzed by plasmid constellation network (placnet), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet*, 10(12):e1004766, 2014.

[80] C. Zhu and J. Yu. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics*, 182(3):875–88, 2009.

[81] S. G. Earle, C. H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon, T. M. Walker, C. C. A. Spencer, Z. Iqbal, D. A. Clifton, K. L. Hopkins, N. Woodford, E. G. Smith, N. Ismail, M. J. Llewelyn, T. E. Peto, D. W. Crook, G. McVean, A. S. Walker, and D. J. Wilson. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*, 1:16041, 2016.

[82] R. B. Cattell. The scree test for the number of factors. *Multivariate Behav Res*, 1(2):245–76, 1966.

[83] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[84] S.; Park S. Kim, D; Kim. Inferences of coordinates in multidimensional scaling by a bootstrapping procedure in r. *Universal Journal of Educational Research*, 3(6):402–406, 2015.

[85] K. J. Shaw, P. N. Rather, R. S. Hare, and G. H. Miller. Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiol Rev*, 57(1):138–63, 1993.

[86] K. Shimizu, T. Kumada, W. C. Hsieh, H. Y. Chung, Y. Chong, R. S. Hare, G. H. Miller, F. J. Sabatelli, and J. Howard. Comparison of aminoglycoside resistance patterns in japan, formosa, and korea, chile, and the united states. *Antimicrob Agents Chemother*, 28(2):282–8, 1985.

[87] G. Cox, P. J. Stogios, A. Savchenko, and G. D. Wright. Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ant(2")-ia. *MBio*, 6(1), 2015.

[88] P. C. Shaw, A. C. Liang, K. M. Kam, and J. M. Ling. Presence of stra-strb gene within a streptomycin-resistance operon in a clinical isolate of shigella flexneri. *Pathology*, 28(4):356–8, 1996.

[89] M. Barlow and B. G. Hall. Experimental prediction of the evolution of cefepime resistance from the cmy-2 ampc beta-lactamase. *Genetics*, 164(1):23–9, 2003.

[90] L. Ma, Y. Ishii, F. Y. Chang, K. Yamaguchi, M. Ho, and L. K. Siu. Ctx-m-14, a plasmid-mediated ctx-m type extended-spectrum beta-lactamase isolated from escherichia coli. *Antimicrob Agents Chemother*, 46(6):1985–8, 2002.

[91] J. Alba, Y. Ishii, K. Thomson, E. S. Moland, and K. Yamaguchi. Kinetics study of kpc-3, a plasmid-encoded class a carbapenem-hydrolyzing beta-lactamase. *Antimicrob Agents Chemother*, 49(11):4760–2, 2005.

[92] W. Ke, C. R. Bethel, J. M. Thomson, R. A. Bonomo, and F. van den Akker. Crystal structure of kpc-2: insights into carbapenemase activity in class a beta-lactamases. *Biochemistry*, 46(19):5732–40, 2007.

[93] B. A. Evans and S. G. Amyes. Oxa beta-lactamases. *Clin Microbiol Rev*, 27(2):241–63, 2014.

Chapter 2

Deep mutational scanning of TEM β-lactamase for visualization of alternate sequence spaces of extended-spectrum resistance

2.1 Foreword and Acknowledgements

The bulk of the experimental laboratory work described in this chapter was performed by a group of my undergraduate students and a junior technician that I trained. I would like to thank them for their hard work (in alphabetical order of last name): Francim Aguilar, Yuki Floyd, Lucy Guan, Allan Stamm, David Winograd.

Additionally, I would like to thank Dr. Cherie Musgrove, Melissa Standley and Dr. Manel Camps for contributing significantly to my training in the wet-lab.

2.2 Introduction

Antibiotic resistance is now a worldwide epidemic, causing extended hospitalization and treatment failures for a range of common bacterial infections [1-4]. Resistance is primarily caused by bacterial enzymes that interact directly with antibiotic drug molecules, rendering them ineffective. For instance, β-lactamases, which comprise one of the largest families of such enzymes, confer resistance via lysis of

the β-lactam ring, a core structure necessary for activity for β-lactam antibiotics (e.g., penicillins, cephalosporins, monobactams, and carbapenems). Resistance to β-lactam antibiotics has steadily risen since the introduction of penicillin to hospitals in 1942. Presently, thousands of β-lactamase variants have been discovered, arising in bacteria through co-evolution between enzyme and substrate (i.e. variants arise from natural mutative processes and the ones with increased activity against specific antibiotics are selected), which is driven by antibiotic exposure. This type of selection more generally applies to many genetic resistance factors and serves as a tailoring process by which resistance enzymes become structurally fitted to have increased activity against specific antibiotics.

Indeed, human action is the predominant force driving both the evolution of resistance and spread of multidrug resistant bacteria [5]. The epidemic is fueled by routine antibiotic uses for human/animal health and food production, and is further intensified by irresponsible uses by patients, uninformed and/or incentivized prescriptions by healthcare providers and misdiagnosis.

Thus, recent research has produced a range of potential alternative treatment methods for bacterial infections. Many investigative efforts have moved away from using antibiotics altogether with immunological therapies, probiotics, and ionic liquids that disrupt bacterial biofilm [6]. Other approaches aim to enhance bacterial susceptibility to existing antibiotics. One such method involves engineering viruses with gene silencing mechanisms for eliminating antibiotic resistance genes [7].

Some methods leverage knowledge of evolutionary mutation pathways to inform antibiotic treatment strategies. In essence, better strategies for slowing or limiting the evolution of resistance can be inferred by studying how resistance arises at the molecular level in antibiotic resistance genes. For example, widely employed antibiotic selection and cycling [8, 9] strategies can be designed rationally, rather than relying on commonly practiced "trial and error" strategies, which are more likely to lead to resistance proliferation. Also, knowledge of evolutionary pathways can facilitate the design of novel antibiotics with desired specificities [10].

Evolutionary pathway studies involve characterizing the effects of multiple mutations in drug target proteins such as β-lactamases. Multiple mutations often participate in epistatic interactions (i.e. have non-additive phenotypic effects), which are shaped by biophysical constrains of protein structure. An improved understanding of how epistasis impacts the translation of genotype to phenotype would enable better treatment strategies and more rational methods for drug design.

During the experimental or clinical evolution of antibiotic resistance sim-

ilar mutational patterns can emerge independently (i.e. in different patients or experimental cultures treated with the same drugs). Information on the impact of individual or multiple mutations can be inferred by observing similar patterns across large sequencing datasets. However, this is often not a simple task. For one, epistatic interactions are derived from a high-dimensional mechanistic framework (i.e. epistasis is intimately tied to protein structure), which are inherently difficult to describe in low-dimensional settings (i.e. 2- or 3-dimensions). Also, random mutational processes introduce noise to sequencing data. For these reasons methods detecting mutation patterns should be robust to noise, or have preprocessing steps for filtering noise and/or reducing data dimensionality. Finally, large datasets are usually required for sufficient statistical power.

Datasets for parallel evolution studies are most efficiently obtained through next-generation sequencing (NGS), for which samples are generally obtained in one of two ways: clinically through the routine sampling of drug treated patients, or experimentally through directed evolution. One emerging application of NGS involves deep sequencing of randomly mutated proteins (amplicon sequencing), subjected to antibiotic selection in bacteria. Such deep mutational scanning studies provide an assessment of the fitness effects of all possible or likely amino acid substitutions across a stretch of protein [11]. Deep mutational scanning can shed light on β-lactamase fitness landscapes, help anticipate evolutionary trajectories, and accelerate the discovery of new resistance conferring mutations.

Here, we perform directed evolution of the TEM β-lactamase under 3rd-generation cephalosporin selection using a deep mutational scanning (DMS) approach, with the aim of exploring new protein sequence subspaces of extended-spectrum resistance that have not been reported previously.

The directed evolution of wild-type TEM-1 under cefotaxime selection is expected to result in a sequence subspace that largely converges around substitution G238S, which confers high-level extended-spectrum resistance, and has been observed widely in clinical and environmental isolates, both independently and in epistasis with other resistance-augmenting mutations.

To provide access to multiple distinct sequence subspaces, we initiate directed evolution from multiple different starting points (using different parental alleles) and reduce the likeliness of convergence between these lineages by leveraging negative epistasis between their initial trajectories.

We describe here, three deliverables: (1) TEM β-lactamase mutant sequence libraries evolved from three independent parental alleles, which can be stud-

ied further to gain a deeper understanding of how negative epistasis impacts protein sequence space divergence under positive selection in a gain-of-function setting, (2) new previously unreported alleles that confer extended-spectrum resistance, and (3) a high-throughput deep mutational scanning procedure used to generate these mutant libraries. We confirm that the sequence landscape contained in our directed evolution libraries spans across multiple distinct protein sequence subspaces, and we verify extended-spectrum resistance for select mutant alleles from each sequence subspace, by reconstructing individual mutant alleles via site-directed mutagenesis and assessing their ability to confer cefotaxime resistance to an *E. coli* host.

2.3 Methods

We evolved three independent lineages from parental allele sequences representing different starting points in the evolution of TEM-1. The three parental alleles corresponded to the wild-type TEM-1 β-lactamase, and two variants containing single amino acid substitutions, R164H and A237T. Both of these single amino acid substitutions have been found to confer a gain-of-function phenotype as shown by decreased susceptibility against cefotaxime, and also have been shown to be incompatible with the G238S substitution which generally dominates the gain-of-function landscape accessible via directed evolution of TEM-1 under cefotaxime selection.

We performed directed evolution of each parental allele using a deep mutational scanning approach. Our approach can be outlined in three steps. First, random mutagenesis is performed via error-prone PCR on a given parental allele, resulting in a mutant sequence library. Second, sequence variants encoding resistance-conferring protein isoforms are selected by expressing individual variants from the mutant sequence library in a susceptible *E. coli* strain grown under cefotaxime selection. Third, long-read amplicon sequencing is performed on the mutant sequence library, both before and after selection, to quantify the relative fitness gain conferred by individual sequence variants under selection.

2.3.1 Generation of plasmid constructs containing parental alleles

A gene containing the wild-type sequence of TEM-1 β-lactamase was initially obtained from the commercially available plasmid pGFPuv. The R164H and A237T alleles were generated by inducing single amino acid substitutions on the wild-type gene by site-directed mutagenesis.

Each parental allele sequence was cloned into a custom plasmid vector pGPSori for storage and phenotype testing. The pGPSori vector is a high-copy plasmid derived from the commercially available pGPS3 vector. This customized vector has an additional multiple cloning site directly downstream of the replication origin, and a Kanamycin resistance marker that can be used as a secondary selective marker for cloning, so as not to affect the ratio of β-lactamase variants in the mutant libraries. Also in pGPSori, the original β-lactamase gene present in pGPS3 is inactivated by truncation.

Each parental allele was inserted into the multiple cloning site downstream of the replication origin in pGPSori, resulting in three parental plasmid constructs named pGPSori-TEM1, pGPSori-TEM1-R164H, and pGPSori-TEM1-A237T.

2.3.2 Random mutagenesis of parental alleles

Each parental allele sequence was mutated by error-prone PCR (epPCR), using the GeneMorph II Random Mutagenesis Kit from Agilent. We designed the epPCR primers for amplification of a 983 kb region of pGPSori containing the parental allele sequence. This region included flanking restriction sites, KpnI and NspI, to be used for cloning. An additional 100-200 base-pairs outside the restriction sites on both ends were included in the amplified region to facilitate size selection of the epPCR amplicon for cloning. The total size of the amplified region was 1273 kb.

The mutation rate of epPCR can be adjusted using the amount of template DNA included in a given PCR reaction. The average number of mutations induced on the amplified DNA segment is inversely proportional to the amount of template DNA included. The GeneMorph II User's Manual offers some guidelines on controlling the mutation rate by adjusting the amount of template DNA; however, these guidelines only provide rough approximations.

We performed several diagnostic rounds of epPCR to determine the experimental conditions required to achieve our desired mutation rate (2 to 3 mutations across the target gene and 4 to 5 mutations across the amplicon). We varied the amount of template DNA in individual epPCR reactions from 600 ng to 900 ng, and counted the number of nucleotide substitutions observed across the length of the target gene (from the start to stop codon) after sequencing. We determined that 700 ng most closely approximates the desired mutation rate.

Error-prone PCR was performed in individual 50 μl reactions, including 700 ng of template DNA (a given parental pGPSori construct), 1 μL of 40 mM

dNTPs, 0.5 μL primer mix (250 ng/μL for each primer), 1 μL of the mutazyme, 5 μL 10x buffer and nuclease-free water. A thermocycler run was executed using the parameters recommended in the GeneMorph II User's Manual.

2.3.3 Preparation of plasmid-borne TEM β-lactamase mutant libraries

We generated a plasmid-borne TEM β-lactamase mutant library corresponding to each parental allele by cloning each set of epPCR amplicons into high-copy expression vectors. Our cloning target vector was a version of pGPSori-TEM-1 called pLA230.2, which encoded a TEM-1 protein that had been inactivated by truncation.

The epPCR amplicons were purified using the Nucleospin PCR clean-up protocol from Macherey-Nagel, then digested alongside pLA230.2 with the Kpn1-HF and Nsp1 restriction enzymes from New England Biolabs. Double restriction digests were performed in 50 uL of reaction volume with 1000 ng of DNA (insert or vector), 1 μL each of Kpn1-HF and Nsp1, 5 μL of 10x CutSmart buffer and nuclease-free water. Following the restriction digests, insert and vector DNA were size-selected and purified on a 0.8% agarose gel. Insert and vector DNA was extracted from the agarose gel matrix using the Macherey-Nagel gel-extraction kit.

Ligation was performed using the Anza T4 Ligase Master Mix. Digested insert and vector DNA were combined in a 2:1 ratio based on molecular weight. Ligation reactions were set up on ice with 5 μL of Anza Master Mix and nuclease-free water in 20 μL reaction volumes. The reactions were removed from ice and incubated at room temperature for 15 minutes to allow for the ligation to take place. Ligated clones were rescued from the ligation reaction by chemical transformation into Top10 *E. coli* hosts, with selection of transformants on LB agar supplemented with 50 μg/ml Kanamycin.

Transformants putatively containing ligated vectors were suspended in LB via plate-washing with 2mL LB. Plasmid DNA was extracted using the PureYield Midiprep Kit from Promega.

The sequence diversity of the resulting plasmid-borne mutant libraries is limited by the efficiency of the cloning. We followed several best practices to maximize the efficiency of cloning. (1) The cloning process is conducted using only high quality DNA, both in purity and concentration (greater than 300 ng/μl). (2) The UV exposure during gel size selection of restriction fragments was limited to less

than one minute, to minimize the potential for DNA damage and sticky-end degradation. (3) A best insert-to-vector ratio was chosen by observing the cloning output from a range of ratios. (4) Digested insert and vector are ligated immediately, or within 48 hours of the restriction digest to minimize sticky-end degradation. (5) The Anza T4 ligase is kept on ice prior to the 15 minute incubation period, or stored at -20 at all times. Pre-chill any apparatus as necessary. (6) All of the DNA from a given ligation reaction is immediately transformed into E. coli hosts. (7) Determine a protocol to maximize transformation efficiency, and total transformation output.

In addition, we included a second amplification step in the cloning procedure as a time and cost-saving measure. We performed high-fidelity amplification of the epPCR amplicons using the Phusion DNA polymerase, and proceeded with cloning using the Phusion amplicons. Using this method, the sequence diversity generated by a single epPCR run can be captured more efficiently during cloning. In essence, the lower ratio of sequence diversity to DNA mass in the Phusion amplicons is sufficient for delivering the maximum sequence diversity that can be captured by a single cloning run producing around 500 to 2000 unique plasmid-borne mutant sequences. Also, the higher DNA mass resulting from Phusion amplification enables parallelization of the cloning procedure. Phusion PCR reactions contained 12.5 μL of Phusion Master Mix, 1.25 μL each of 10 uM forward and reverse primers (designed to amplify the entire epPCR product), 10 ng of epPCR product and nuclease-free water to achieve 25 μL reaction volume.

2.3.4 Enrichment for catalytically active protein isoforms

Prior to selection with cefotaxime, we first performed a preliminary selection to enrich our mutant libraries for sequence variants encoding catalytically active β-lactamases. This pre-selection was performed using a low dosage of carbenicillin to enforce a weak selection for β-lactamase variants that have retained the ability to cleave a 1st-generation cephalosporin. Since activity against 1st-generation cephalosporins is a structural prerequisite for extended-spectrum activity, the set of variants selected in this manner should include variants that have gained the ability to cleave extended-spectrum β-lactams. By performing this pre-selection step, we increase the throughput of relevant sequence diversity for the subsequent selection and sequencing steps. We also remove any undigested cloning target vectors (pLA203.2) that may have remained after size selection.

Pre-selection was performed by expressing the mutant libraries in Top10 *E.*

coli hosts. Plasmid DNA from each mutant library was transformed into chemically competent Top10 cells, and plated on LB agar supplemented with 33 μg/ml carbenicillin. Plasmid DNA was extracted from transformant colonies remaining after overnight growth, using the PureYield Midiprep Kit from Promega.

2.3.5 Selection with cefotaxime in E. coli hosts

Each mutant library (corresponding to one of three parental alleles: wild-type TEM-1, R164H, or A237T) was subjected to selection with cefotaxime in *E. coli* hosts, at dosages that provided 90% killing of the hosts. As a control, the R164H and A237T libraries were additionally selected at dosages registering a minimum noticeable killing to first verify that our dosages were in an appropriate range to impact the sequence diversity (spectrum of mutants generated) of selected mutants. Cefotaxime selection was performed using the same general procedure as the pre-selection step.

In total, we generated 8 plasmid-borne mutant libraries for sequencing, with names corresponding to the corresponding parental plasmids: (1) unselected pGPSori-TEM-1, (2) unselected pGPSori-TEM-1-R164H, (3) unselected pGPSori-TEM-1-A237T, (4) high dosage selection pGPSori-TEM-1, (5) high dosage selection pGPSori-TEM-1-R164H, (6) high dosage selection pGPSori-TEM-1-A237T, (7) low dosage selection pGPSori-TEM-1-R164H, (8) low dosage selection pGPSori-TEM-1-A237T.

We performed a preliminary round of sequencing of these eight libraries to verify that we had successfully obtained the desired mutation load from error-prone PCR, and the had captured adequate sequence diversity at each level of selection. Sanger sequencing was performed to amplify the β-lactamase encoding regions in 16 randomly chosen mutants from each library.

2.3.6 Long-read amplicon sequencing

Long-read amplicon sequencing was performed using the PacBio Sequel II system, via a sequencing service provided by GeneWiz LLC.

Amplicons from each library were generated by PCR using the high-fidelity Phusion polymerase, with 6-nucleotide inline barcodes added as forward primer overhangs. To achieve base balance across the 8 barcodes, we selected 8 of the barcodes from Illumina's standard 6nt barcode sequences. PCR products containing barcoded amplicons from each library were purified using the PCR clean-up kit from

Macherey-Nagel, and mixed in equimolar ratios in a single 1.5mL Eppendorf tube.

Prior to PacBio sequencing library preparation, AMPure bead clean-up was performed prior to reduce the presence of potential contaminants and to reconstruct damaged amplicon DNA.

2.4 References

1. Spellberg, B., et al., The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. Clin Infect Dis, 2008. 46(2): p. 155-64.

2. Oxford, J. and R. Kozlov, Antibiotic resistance–a call to arms for primary health-care providers. Int J Clin Pract Suppl, 2013(180): p. 1-3.

3. World Health Organization., The evolving threat of antimicrobial resistance: options for action. 2012, Geneva, Switzerland: World Health Organization. ix, 119 p.

4. World Health Organization., Antimicrobial resistance: global report on surveillance. 2014, Geneva, Switzerland: World Health Organization. 256.

5. Palumbi, S.R., Humans as the world's greatest evolutionary force. Science, 2001. 293(5536): p. 1786-90.

6. Zakrewsky, M., et al., Ionic liquids as a class of materials for transdermal delivery and pathogen neutralization. Proc Natl Acad Sci U S A, 2014. 111(37): p. 13313-8.

7. Yosef, I., et al., Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. Proc Natl Acad Sci U S A, 2015. 112(23): p. 7267-72.

8. Abel zur Wiesch, P., et al., Cycling empirical antibiotic therapy in hospitals: meta-analysis and models. PLoS Pathog, 2014. 10(6): p. e1004225.

9. Goulart, C.P., et al., Designing antibiotic cycling strategies by determining and understanding local adaptive landscapes. PLoS One, 2013. 8(2): p. e56040.

10. Planson, A.G., et al., Engineering antibiotic production and overcoming bacterial resistance. Biotechnol J, 2011. 6(7): p. 812-25.

11. Fowler, D.M. and S. Fields, Deep mutational scanning: a new style of protein science. Nat Methods, 2014. 11(8): p. 801-7.

Chapter 3

Transposable element islands in the inbred invasive ant *Cardiocondyla obscurior* facilitate adaptation to novel environments

3.1 Foreword and Acknowledgements

This chapter describes collaborative work that was published in Nature Communications in 2014 (Schrader, L. et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat. Commun. 5:5495 doi: 10.1038/ncomms6495 (2014)). In the following, the main text of the manuscript published in Nature Communications is provided with references to Supplementary Information, which can be found at https://www.nature.com/articles/ncomms6495.

My main contribution to this work consists of a simple repeat and transposable element (TE) detection pipeline that was used to annotate the repeat landscape of eight ant genomes (*Acromyrmex echinatior* (Aech), *Atta cephalotes* (Acep), *Solenopsis invicta* (Sinv), *Linepithema humile* (Lhum), *Pogonomyrmex barbatus* (Pbar), *Harpegnathos saltator* (Hsal), *Camponotus floridanus* (Cflo)), *Cardiocondyla obscurior* (Cobs)) the parasitic wasp *Nasonia vitripennis* (Nvit) and the honeybee *Apis mellifera* (Amel). The data obtained through this pipeline enabled the discovery and study of TE islands in context of the invasive and inbreeding ant

Cardiocondyla obscurior.

Here, I would like to thank and recognize the efforts of my co-authors, especially Dr. Lukas Schrader and Dr. Jan Oettler, who spearheaded the project and writing of the manuscript. They are the main contributors to this work. And also Dr. Christopher D. Smith who advised me during the writing of the early version of the TE detection pipeline used in this study.

3.2 Introduction

Depletion of genetic variation is detrimental to species evolution and adaptation [1]. Low genetic and phenotypic variation is common in founder populations, where only one or a few genotypes are isolated from a source population. Under such conditions, reduced effective population size (Ne) should decrease selection efficiency and increase genetic drift, resulting in only weak selection against mildly deleterious alleles which can thus accumulate [2]. These effects should be even stronger in inbreeding species [3] and taxa with generally low Ne such as social insects [4]. Despite these constraints on adaptive evolution, many inbred or selfing species thrive and are able to invade novel habitats. This raises the question of how genetic variation as the raw material for adaptation is generated in such systems.

Single-nucleotide substitutions are an important factor in adaptation [5] and species diversification [6,7]. However, other structural and regulatory units, such as transposable elements (TEs) and epigenetic modifications, may act as drivers in adaptation and evolution [8]. TEs play a particularly vital role in genome evolution [9] and recurringly generate adaptive phenotypes [10-13] primarily through (retro-)transposition [14], and secondarily through ectopic recombination and aberrant transposition [15].

The invasive, inbreeding ant *Cardiocondyla obscurior* (Fig. 1) provides a suitable model to study how species adapt to novel habitats in spite of constraints imposed by invasion history, life history or both. Originally from Southeast Asia, *C. obscurior* has established populations in warm climates around the globe from founder populations that presumably consisted of only one or a few inbred colonies, each with a few reproductive queens and several dozen sterile workers. In this species, related wingless males and females (queens) mate within the colony, after which queens leave the colony with a group of workers to find a new nest nearby. While greatly reducing the extent of gene flow between colonies, this behaviour enables sexual reproduction within the same colony and allows single founder colonies to

Figure 3.1: Two workers of C. obscurior and the remains of a fly. Hidden in small cavities of plants, the inconspicuous colonies of this species are frequently introduced to new habitats by global commerce. In spite of strong genetic bottlenecks, even single colonies with few reproductive individuals suffice to establish stable populations.

rapidly colonize novel habitats. At the same time, the combination of prolonged inbreeding with severe genetic bottlenecks strongly reduces Ne in this species. Under such conditions, genetic drift is predicted to drastically deplete genetic variation, thus leaving little for selection to act on.

Here we explore the genomes of *C. obscurior* from two invasive populations (Brazil BR and Japan JP) to identify signatures of divergence on a genomic level and to determine how the species can rapidly adapt to different habitats. We find clear phenotypic differences between the populations and strong correlation between accumulations of TEs ('TE islands') and genetic variation. Our results suggest that TE islands might function as spring wells for genetic diversification in founder populations of this invasive species. The distinct organization of TE islands, their gene composition and their regulation by the genome adds compelling evidence for the role of TEs as players in differentiation, adaptation and speciation.

3.3 Results

3.3.1 Phenotypic differences between BR and JP lineages

Colonies from the two populations contained similar numbers of workers (Mann–Whitney U-test=778.5, $Z = -0.634$, P=0.526; BR: median=28, quartiles 21.75 and 51.25, n=27 colonies; JP: median=29, quartiles 16 and 47, n=64), but queen number was higher in Japan (Mann–Whitney U-test=501, $Z = -3.084$, $P < 0.003$; BR: 5 queens, quartiles 3, 8; JP: median=10, quartiles 4 and 19). Body sizes of queens and workers from BR were significantly smaller than in JP individuals, yet wingless males did not differ in any of the measured characters.

In ants, cuticular chemical compounds play a particular prominent role in kin recognition, which is crucial for species integrity but on a deeper level also a requirement for the maintenance of altruism [16]. Analysis of cuticular compound extracts from BR and JP workers showed that compound composition differed significantly between the two lineages (multivariate analysis of variance: df=2, F=10.33, R2=0.39, P¡0.001) and samples were classified correctly according to population of origin in 83.3% of cases (Supplementary Table 1; Supplementary Fig. 1).

The lineages also differed in behaviour, with BR colonies being significantly more aggressive towards both workers and queens from their own lineage, while JP colonies more readily accepted JP workers and queens (PWorkers $JP \times JP$ versus $BR \times BR = 0.000296$, PQueens $JP \times JP$ versus $BR \times BR = 7.98e-07$, Supplementary Fig. 2). Confronted with individuals from the other lineage, BR colonies were as aggressive as in within-population encounters (PWorkers $BR \times JP$ versus $BR \times BR = 0.39$, PQueens $BR \times JP$ versus $BR \times BR = 0.94$), while JP colonies were again significantly less aggressive (PWorkers $JP \times BR$ versus $BR \times BR = 0.000131$, PQueens $BR \times JP$ versus $BR \times BR = 1.23e-07$). Testing discrimination against workers of another ant species, *Wasmannia auropunctata*, evoked similarly high aggressive responses in both lineages, suggesting that the BR and JP populations do not generally differ in their aggressive potential.

3.3.2 The *C. obscurior* genome is compact and rich in class I TEs

Using MSR-CA version 1.4, we produced a 187.5-Mb draft reference genome based on paired-end sequencing of several hundred diploid females (454 Titanium FLX sequencing) and a 200-bp library made from five haploid males (Illumina HiSeq2000; Supplementary Table 2), all coming from a single Brazilian colony. Au-

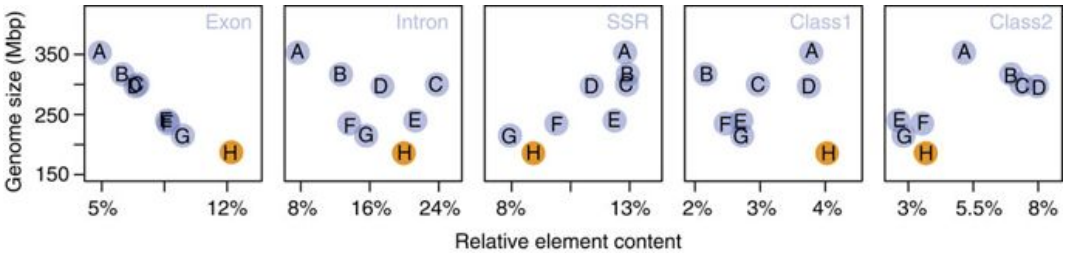

Figure 3.2: Assembly size in Mbp plotted against the relative proportion of exons, introns and different repetitive elements. The analysed genomes show a negative correlation between relative exon but not intron content. Genome size is positively correlated with relative short simple repeat but not class I and II TE content. A, *S. invicta*; B, *A. cephalotes*; C, *A. echinatior*; D, *H. saltator*; E, *C. floridanus*; F, *P. barbatus*; G, *L. humile*; H, *C. obscurior*.

tomatic gene annotation using MAKER version 2.20 (ref. 17) was supported by 454 RNAseq data of a normalized library made from a pool of all castes and developmental stages. We filtered the assembly for prokaryotic scaffolds and reduced the initial 11,084 scaffolds to 1,854 scaffolds, containing all gene models and a total of 94.8% (177.9 Mb) of the assembled sequence. The genome can be accessed under antgenomes.org/ and hymenopteragenome.org.

The final gene set contains 17,552 genes, of which 9,552 genes have a known protein domain as detected by IPRScan (www.ebi.ac.uk/interpro/), and falls within the range of recent estimates for eight other sequenced ant species [18-26]. Of all genes, 72.5% have an annotation edit distance of less than 0.5, which is consistent with a well-annotated genome [27] (Supplementary Table 3).

The *C. obscurior* genome is the smallest so far sequenced ant genome [18-26]. Although there is no physical genome size estimate for *C. obscurior*, assembled sequences and physical estimates are tightly correlated in seven ant genomes (LM in R: R2=0.73, F1, 5=13.7, P=0.014, from [28]), suggesting that *C. obscurior* has the smallest genome reported so far for an ant species [29]. Overall, the draft genome size of the analysed sequenced ants is negatively correlated to relative exon content (GLM in R: $df = 6$, $F = 150.55$, $P < 0.001$) but not to relative intron content ($df = 5$, $F = 0.65$, $P = 0.460$; Fig. 2), indicative of stabilizing selection on coding sequence. In contrast, intron size distribution is diverse between ant genomes and is not correlated with genome size (Supplementary Fig. 3; Supplementary Table 4).

We used a custom pipeline (see Supplementary Information) to identify simple repeats, class I retrotransposons and class II DNA transposons in *C. obscurior*, seven ant genomes (*Acromyrmex echinatior* (Aech), *Atta cephalotes* (Acep), *Solenopsis invicta* (Sinv), *Linepithema humile* (Lhum), *Pogonomyrmex barbatus*

(Pbar), *Harpegnathos saltator* (Hsal), *Camponotus floridanus* (Cflo)), the parasitic wasp *Nasonia vitripennis* (Nvit) and the honeybee *Apis mellifera* (Amel). Across the analysed ants, genome size is significantly correlated with relative simple repeat content (lm, R2=0.66, F=11.83, P=0.014; Fig. 2) but not with class I and class II TE content. However, it appears that the larger genomes contain more relative class II sequence. Relative class I retrotransposon content was highest in *C. obscurior* (7.6 Mb, 4.31%, Supplementary Fig. 4) and in particular, many class I non-LTR retrotransposons (for example, 14 types of LINEs) and several types of LTR transposons (Ngaro, Gypsy, DIRS and ERV2), TIR elements (for example, hAT, MuDR, P) and Helitrons are more abundant in *C. obscurior* (Supplementary Table 5).

3.3.3 Genomic signatures of an inbred lifestyle

On the basis of TE content calculations for 1 and 200 kb sliding windows, we identified 18 isolated 'TE islands' located in 'LDR' (low-density regions) in the *C. obscurior* genome. These TE islands were defined as containing TE accumulations in the 95-100% quantile within scaffolds over 200 kb (87 scaffolds, representing 96.02% or 170.8 Mb of the assembly). In total, TE islands cover 12.78 Mb of sequence (7.18% of total sequence) and range between 0.19 and 1.46 Mb in size. The TE islands contain 27.54% (4.92 Mb) of the assembly-wide TE sequence (17.87 Mb), 6.6% of all genes (1,160), and have reduced exon content (TE islands 87.0 exon bp kb-1, LDRs 124.5 exon bp kb-1). Note that some larger scaffolds contain more than one TE island.

Retroelements of the superfamilies BEL/Pao, DIRS, LOA/Loa, Ngaro, R1/R2 and RTE as well as DNA transposons of the superfamilies Academ, Kolobok-Hydra, Maverick, Merlin, on and TcMar-Mariner/-Tc1 populate TE islands with significantly higher copy numbers than other elements (Fisher's exact test, false discovery rate < 0.05, Fig. 3, Supplementary Table 6). Furthermore, both class I and class II elements show a length polymorphism, with elements in TE islands being significantly longer compared with elements in LDRs (U-tests, $W = 109089018$, $P < 2e-16$ for class I and $W = 152340067$, $P < 2e-16$ for class II, Fig. 4a, Supplementary Fig. 5).

We also assessed the genome-wide TE distributions for seven published ant genomes, Amel v4.5 and Nvit v2.0 (Fig. 5). The smaller ant genomes (Pbar, Lhum and Cflo) and Amel are similar in TE sequence distribution. In contrast, the larger genomes (Aech, Acep, Sinv and Hsal) are more variable, have higher median

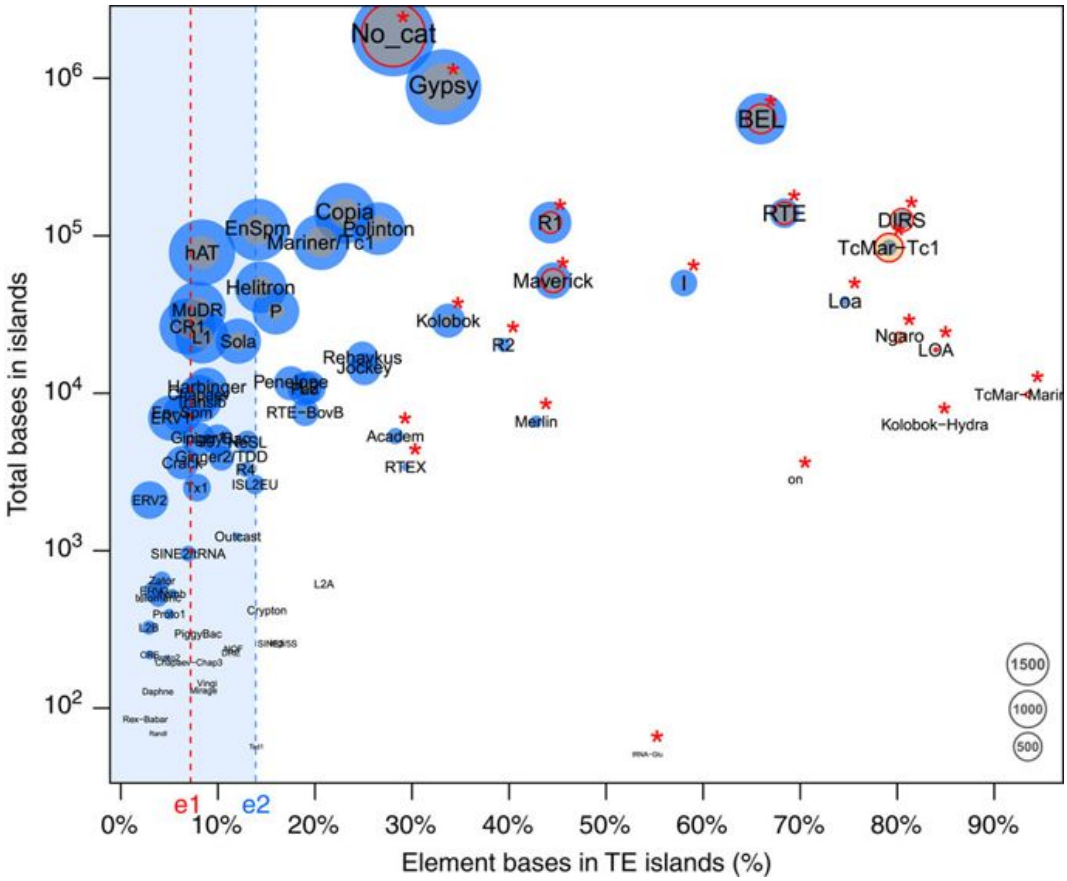

Figure 3.3: The proportion of bases annotated in TE islands in C. obscurior against the log-scaled total base count in TE islands for each TE superfamily. Point size is relative to the copy number of the respective element found in TE islands (orange) and in LDRs (blue). Red circles indicate superfamilies with significantly higher frequency in TE islands than other superfamilies. Superfamilies with a significantly higher base count in TE islands are denoted by a red asterisk. e1: Percentage of the genome contained in TE islands (7.18%), e2: median across all types of TEs (13.89%).

TE content and a much broader and tailed TE frequency distribution with longer stretches of high or low TE content. The genome of *C. obscurior* is distinct from the other ant genomes, with low TE content in LDRs but exceptional clustering with high TE densities in TE islands. The genome of the inbred wasp *N. vitripennis* contains regions with up to 60% TE content that are surrounded by LDRs containing much less TE sequence (10%), resembling the pattern observed in *C. obscurior*.

3.3.4 TE islands diverge faster than LDRs in the two populations

We mapped 140 Gb of genomic DNA Illumina reads (60 × coverage for each population) from pools of 30 (BR) and 26 (JP) male pupae, respectively, against the reference genome (BWA; bio-bwa.sourceforge.net) and analysed the local coverage ratio to detect genetic divergence. Deviations from the mean coverage ratio

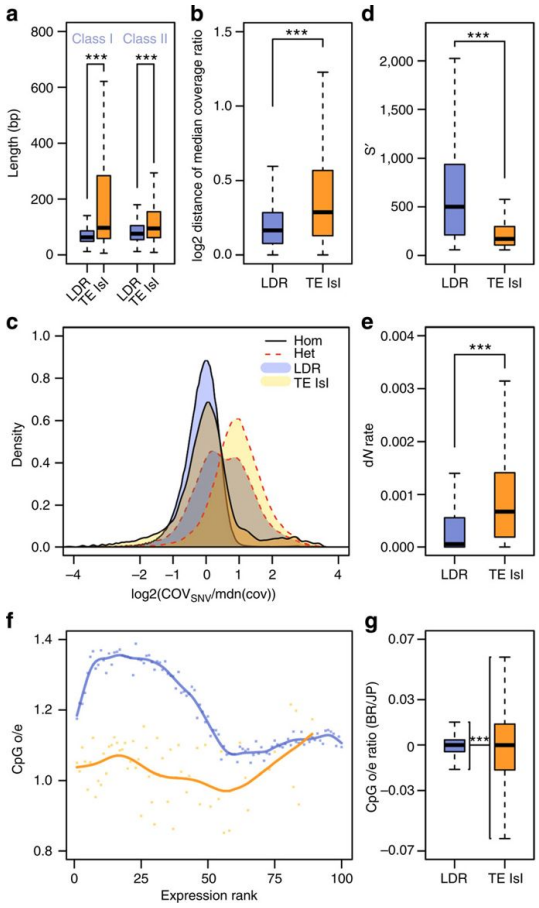

Figure 3.4: Quantitative measures on the divergence of TE islands and LDRs. (a) Length polymorphism for Class I and Class II TEs in LDRs (blue) and TE islands (orange). U-tests, nLDR=54,950, nTE=6,466 for class I and nLDR=59,054, nTE=6,813 for class II. (b) Deviations from the median coverage ratio calculated for 1 kb windows in LDRs (blue) and TE islands (orange). U-test, nLDR=157,296; nTE=12,165. (c) Log2-scaled density plots of the coverage for all homozygous (solid black lines) and heterozygous SNV (dotted red lines) calls divided by the median coverage (orange, calls within TE islands; blue, calls in LDRs). Coverage at homozygous calls is not different from the median overall coverage, neither in TE islands nor in LDRs. The shift for heterozygous SNV calls within TE islands shows that most calls result from diverging duplicated loci. The bimodal distribution for heterozygous calls in other genomic regions suggests two distinct populations of SNV calls, that is, true heterozygous loci (first peak) and diverging sequence in duplicated loci (second peak). (d) Bit scores for genes in LDRs (blue) and TE islands (orange) retrieved by BLASTx against annotated proteins from seven ant genomes. U-test, nLDR=12,065; nTE=902. (e) Rates of non-synonymous substitutions (calculated as $dN/(dN+dS)$) in LDR (blue) and TE island genes (orange). U-test, nLDR=6,806; nTE=423. (f) Exon-wide CpG o/e values were plotted against the expression rank from 0 (least expressed) to 100 (most expressed) genes for LDRs (blue) and TE islands (orange). (g) Calculated ratios (BR/JP) for exon CpG o/e values in LDRs (blue) and TE islands (orange). F-test, $nLDR = 16,379$; $nTE = 1,159$. ($^{***}P < 0.0001$, boxplots show the median, interquartile ranges (IQR) and 1.5 IQR.).

(Fig. 6) are in part caused by sequence deletions, insertions and duplications [30]. Such variations are particularly frequent in TE islands (Figs 4b and 6), suggesting accelerated divergence within islands (median deviation from mean coverage ratio:

0.288 in TE Islands, 0.163 in LDRs; U-test, $W = 640300902$; $P < 2e - 16$).

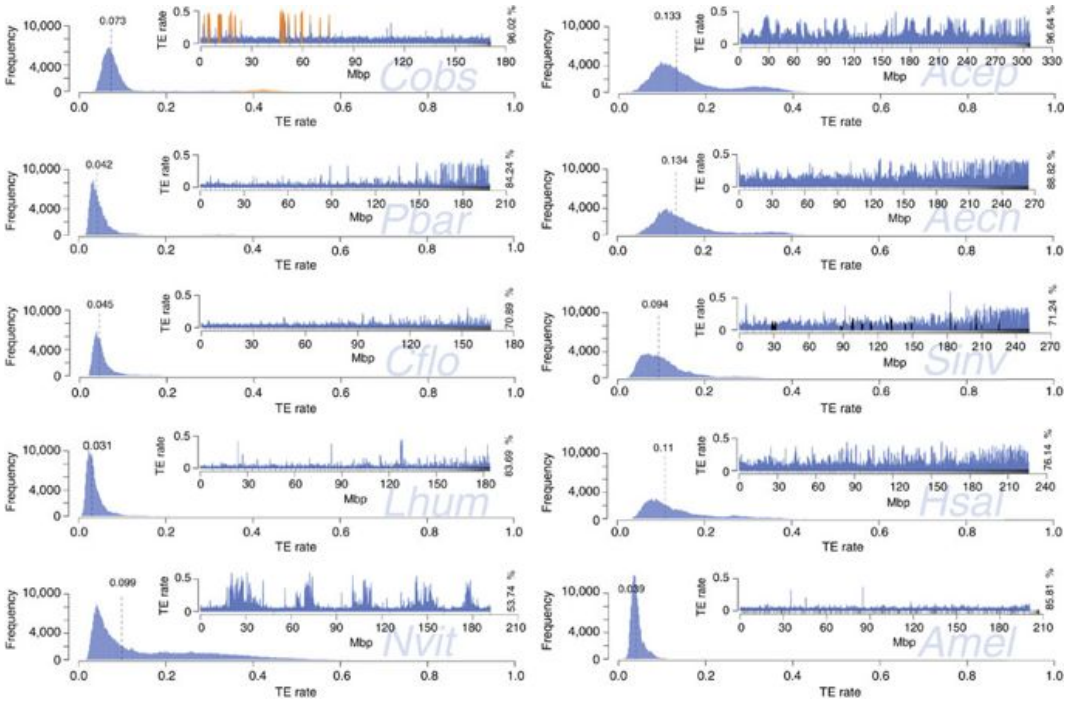

Figure 3.5: Frequency and distribution (insert plots) of TE content in 200 kb windows. Frequency plots: dashed lines denote median TE content. Distribution plots: different proportions of total draft genome sequence were analysed (in %), depending on assembly quality. Scaffolds are sorted by size, small upward tick marks indicate scaffold boundaries. For *C. obscurior*, regions defined as TE islands are coloured in orange. For *S. invicta*, scaffolds mapping to a non-recombining chromosomal inversion [73] are shown in black. For *A. mellifera*, scaffolds were sorted according to linkage group.

We retrieved SNV (single-nucleotide variants) calls using consensus calls from samtools (samtools.sourceforge.net) and the GATK (broadinstitute.org/gatk/). Although TE islands only comprise 7.18% of the genome, they combine 15.59% (86,236 of 553,052) of all SNV calls. Given that we sequenced haploid males from highly inbred lineages, heterozygous SNVs should be rare. A large fraction of heterozygous SNVs in both lineages are within TE islands (62.95% of 62,879 in BR, 50.52% of 98,353 in JP), while rates of homozygous calls (Fig. 6) are not increased (11.88% of 16,277 in BR, 6.91% of 445,316 in JP). High numbers of false positive heterozygous SNVs calls can arise in duplicated regions that collapsed into a single locus due to misassemblies [31]. Accordingly, such SNVs can be identified by a twofold increase in coverage and in fact mark diverging duplicated loci within the same lineage (Fig. 4c).

Genes in TE islands should also show signatures of accelerated divergence from orthologues if overall sequence evolution is increased in these regions. Indeed,

BLASTp searches against seven ant proteomes produced significantly lower bit scores for genes within TE islands when compared with genes in LDRs (Fig. 4d, U-test, $W = 120460260$, $P < 2e-16$). In accordance, SNV annotation revealed higher rates of non-synonymous substitutions between the BR and JP lineage in TE island genes (Fig. 4e, U-test, $W = 923754$, $P < 2e-16$). Surprisingly, however, on average, TE island genes contained less synonymous SNVs than LDR genes (LDR 0.67 kb-1, TE island 0.42 kb-1, U-test, $W = 10743397$, $P < 2e-16$).

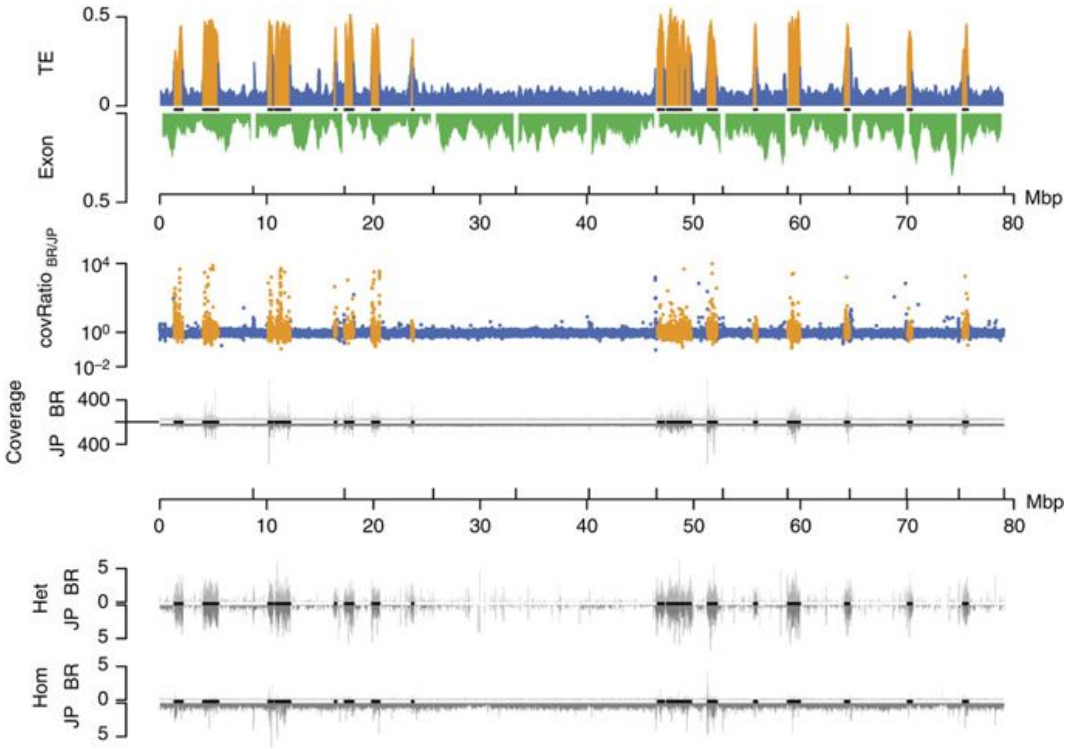

Figure 3.6: Genomic divergence and subgenomic structure of the 12 largest C. obscurior genome scaffolds (including all 18 TE islands). High TE content in TE islands correlates with deviations from the average coverage ratio, very high absolute coverage in both lineages and high numbers of SNV calls. First track: relative TE (blue and orange within TE islands) and exon content (green) per 200 kb. Second track: coverage ratio BR/JP (blue and orange within TE islands). Third track: absolute coverage for BR (top) and JP (bottom). Fourth track: heterozygous SNV calls per kb in BR (top) and JP (bottom) relative to the reference genome. Fifth track: homozygous SNV calls per kb in BR (top) and JP (bottom) relative to the reference genome. Black lines on x axes indicate localization of TE islands.

3.3.5 Copy number variation within and between TE islands

We inspected 512 candidate loci (155 in TE islands) of 1 kb length by plotting the coverage of each lineage relative to SNVs, genes, and TEs at the respective position, to find genes potentially affected by deletion or copy number variation

events and compiled a list of 89 candidate genes (Supplementary Table 7). Experimental proof-of-principle was conducted by PCR and Sanger sequencing for two deletion candidates (Cobs_13563 and Cobs_01070) and by real-time quantitative PCR for four duplication candidates (Cobs_13806, Cobs_17872, Cobs_13486, and Cobs_16853) (Supplementary Fig. 7). A majority of these genes are located in TE islands (61.8%) and 34 genes show at least weak expression in BR individuals in RNAseq data (see below). The affected genes play roles in processes that may be crucial during invasion of novel habitats, such as chemical perception, learning and insecticide resistance. In particular, four different odorant/gustatory receptor genes show signs of either multiple exon (Cobs_05921, Cobs_13418, Cobs_14265) or whole-gene duplication (Cobs_17892). A gene likely involved in olfactory learning, Cobs_13711, a homologue to pst [32], also shows signs of duplication. Three genes homologous to fatty acid synthase (FAS) genes, a key step in cuticular odour production, contain partial deletions (Cobs_16510, Cobs_14262) or duplications (Cobs_15866). Furthermore, we found differences in genes associated with insecticide response (Cobs_00487, a homologue of nAChRα6 (FBgn0032151) (ref. [33]) and Cobs_17834, coding for a homologue to Cyp4c1 (EFN70878.1) (ref. [34]). Other key genes affected are associated with circadian rhythm (Cobs_17789, homologue to per (FBgn0003068)), caste determination (Cobs_01070, with homology to Mrjp1 (gi406090) (ref. [35]), development (Cobs_17755, coding for a homologue of VgR (Q6X0I2.1) (ref. [36]) and aging (Cobs_14758, with homology to Mth2 (FBgn0045637) (ref. [37]).

De novo assembly of 23M Illumina paired-end reads from the JP lineage that could not be mapped to the BR reference genome resulted in 17 contigs after filtering with highly significant BLASTx hits against proteins of other ants, suggesting that these conserved sequences were lost in the BR lineage instead of being gained in the JP lineage. According to functional annotation, among others these contigs code for homologues involved in development (Vitellogenin-like (XP_003689693)) [38], cellular trafficking (Sorting nexin-25 (EGI65030)) [39], immune response (Protein Toll (EGI66069)) [38] and neuronal organization (Peripheral-type benzodiazepine receptor-associated protein 1 (EFN68490)) [40] (Supplementary Table 8).

3.3.6 Gene composition and regulation of TE islands

Increased TE activity may incur costs to fitness by disrupting gene function. A two-tailed Gene Ontology (GO) enrichment analysis revealed that 59 GO terms associated with conserved processes (for example, cytoskeleton organization, ATP

binding, organ morphogenesis) are under-represented in TE islands, while 18 GO terms are enriched (Supplementary Tables 9 and 10). Four of the over-represented terms relate to olfactory receptors (ORs; GO:0004984, GO:0005549, GO:0050911, GO:0007187) and two terms relate to FAS genes (GO:0005835, GO:0016297). The remaining 12 terms most likely relate to TE-derived genes.

Gene body CpG depletion as a result of increased CpG to TpG conversion due to cytosine methylation is a measure for germline methylation (that is, epigenetic regulation) in past generations. In TE island genes, the exon-wide median observed/expected (o/e) CpG ratio is significantly lower than in other genes (t-test, TE island genes: 1.05, LDR genes: 1.20, $P < 1e-16$). However, both sets of genes show strikingly different correlations of expression and o/e CpG values (Fig. 4f). For LDR genes, o/e CpG values are high in moderately expressed genes and low in highly expressed genes. In contrast, in TE islands, weakly to moderately expressed genes contain less CpG dinucleotides, while highly expressed genes have higher o/e CpG values. To further identify traces of differential regulation of TE islands, we compared the exon o/e CpG values between the lineages by calculating BR/JP ratios for each exon's o/e CpG values and found higher variance in BR/JP ratios in TE islands than in LDRs (Fig. 4g, F-test, $F = 0.136$, $P < 2e-16$, ratio of variances=0.136).

Finally, to assess whether gene expression levels differed between LDRs and TE islands, we generated 14 and 17 Gb transcriptomic RNAseq data of seven queens and seven queen-destined larvae (third larval stage), respectively, from the BR lineage. We estimated mean normalized expression values for each gene using DESeq2, revealing that expression in TE islands was much lower than in LDRs (median expression of all LDR genes=25.45; in TE islands: 0.49; U-test, $W = 14461310$, $P < 2e-16$). While larvae and adult queens did not differ in the expression of LDR genes (median expression in queens=21.16; in larvae=23, 72; U-test, $W = 133301709$, $P = 0.221$), TE island genes were more expressed in adult queens (median expression in queens=0.84; in larvae=0; $W = 1031038$, $P < 2e-16$; Fig. 7, see Supplementary Fig. 6 for details on differential expression between queen and larvae).

3.4 Discussion

C. obscurior is a textbook example for successful biological invasion. Its small size allows for interspecific avoidance, it can rapidly establish colonies in dis-

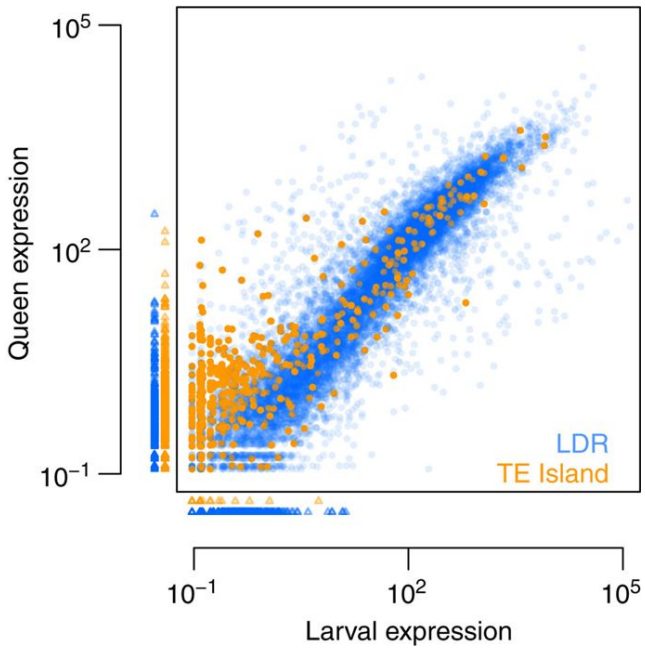

Figure 3.7: Mean normalized expression in third instar queen larvae and mated adult queens for all Cobs1.4 genes. Small triangles indicate genes with no expression in queens (plotted below the x axis) or larvae (plotted left to the y axis). Ninety-five TE island genes and 1,382 LDR genes were not expressed at all (orange, TE island genes; blue, LDR genes).

turbed habitats, and multiple generations per year allow for fast adaptation. While variation in CHCs and body size between the populations point to adaptations to different environments, higher queen number in the JP lineage is likely correlated with reduced intraspecific aggression.

The small genome of *C. obscurior* differs markedly from the other analysed ant genomes in TE distribution and overabundance of several class I subclasses. Importantly, the genome contains low frequencies of TEs in LDRs but well-defined islands with high densities of TEs. In these islands, TEs are on average longer than in LDRs, suggesting overall higher TE activity [41]. Differences in mutation rates and sequence divergence between LDRs and TE islands reveal distinct evolutionary dynamics acting within the *C. obscurior* genome. Moreover, in TE islands, key genes are removed and the majority of genes is less expressed in larvae than adult queens. The non-random distribution of TEs suggests that intragenomic differences in selection efficiency against TEs may have further supported the formation of such locally confined TE accumulations.

Inbreeding can facilitate the accumulation of TEs3 and repeated exposure to stress induced by novel environmental conditions can further amplify TE proliferation [42]. Small Ne is expected to increase the effects of genetic drift and in

turn reduce selection efficiency against mildly deleterious mutations [2]. Under such conditions, local accumulations of TEs might have formed in genomic regions under relaxed selection. Similarly, a reduction in Ne in inbred Drosophila leads to a shift in the equilibrium between TE proliferation and purifying selection against TEs, thus allowing TEs to accumulate [43].

How can we explain extensive proliferation and diversification of TEs within islands, but purifying selection against TEs in LDRs? Coalescent effective population size of a genomic region is positively correlated with its recombination frequency and thus the local efficiency of selection and mutation rate [11]. The initial foundation of TE islands could hence be facilitated in genomic regions with low recombination frequency, providing a refugium of relaxed selection for TE insertions. Indeed, elevated rates of non-synonymous substitutions suggest relaxed selection on TE island genes. Increased frequency of DNA repair processes as a consequence of higher DNA transposition frequencies in TE islands should lead to more errors in DNA replication and double strand break repair [44] in comparison with LDRs. Large-scale mutations on the other hand, such as exon or gene duplications/deletions or gene shuffling, can directly be introduced during TE transposition [45]. TE islands may frequently produce genetic novelty and eventually, by chance, but despite high stochastic drift, adaptive phenotypes, corroborating the view of TEs as genetic innovators.

The list of genes affected by duplications or deletions contains a number of candidates that might be key to the divergence of the lineages. For example, differences in homologues to genes involved in larval development (for example, Mrjp1) might explain body-size differences. Two other candidates, Cobs_00487 and Cobs_17834, show homology to genes that are involved in pesticide resistance against Chlorpyrifos and Imidacloprid (nAChRα6) and Deltamethrin (Cyp4c) in different invertebrate species [46-49]. Imidacloprid treatment of gall wasp infested Erythrina variegate coral trees of the Japan habitat occurred at least once the year before collection of the colonies in 2010 (personal communication S. Mikheyev). In the Brazil habitat, Chlorpyrifos, Deltamethrin and the organophosphate Monocrotophos have routinely been used over the last 10 years (personal communication J.H.C. Delabie).

Furthermore, several within-island genes involved in the production (FAS50) and perception (ORs) of chemical cues contained deletions or duplications in one of the lineages. These results suggest that variation in FAS genes may be responsible for diverging CHC profiles in C. obscurior [51], while variation in OR

genes affects olfactory perception. Chemosensory neurons express highly sensitive ORs52, which are particularly diverse [53] and under strong selection in ants [54]. Gene loss and duplication in the OR gene family has been significantly frequent [55] and differences are assumed to be shaped by adaptive processes in response to a species' ecological niche [56,57]. Intriguingly, the diversification of OR genes is thought to be largely caused by gene duplications and interchromosomal transposition [58], two mechanisms known to be by-products of TE activity. While the distinct patterns of kin recognition and aggressive behaviour in the two lineages of *C. obscurior* may in part be explained by TE-mediated variation in these genes, they also suggest lineage-specific dynamics of the interaction of phenotype and genome evolution. Reduced aggression between colonies in the JP lineage should promote gene flow by exchange of reproductives and thus increase Ne, heterozygosity, and the efficiency of sexual recombination, facilitating the spread of novel arising genotypes. Our findings contrast the view of reduced aggression between colonies of invasive ants [59], but so far it is unclear whether lineage-specific differences are caused by variation in perception or downstream neuronal processes.

Mechanisms controlling TEs are as old as prokaryotes [9] and in fact most TEs are epigenetically silenced [45,60], through either methylation, histone modifications [61] or RNAi [62]. Even though many genes in TE islands are expressed, the overall expression is significantly lower than in LDRs. In line with previous correlations on methylation and expression in eusocial insects [63,64], o/e CpG ratios in *C. obscurior* LDR genes are negatively correlated with expression. However, TE island genes do not follow this trend, in that they are weakly expressed while having low o/e CpG rates. Proximity to TEs can increase gene body methylation [65], which could explain stronger methylation of TE island genes and thus CpG depletion. Also, relaxed selection in island genes should in general increase fixation frequency of base mutations, including CpG to TpG conversions thus depleting CpG content. Gene expression differences in TE island genes between larvae and adult queens suggest stronger regulation of these potentially disruptive genes during the sensitive developmental phase. Finally, key regulatory genes are under-represented in TE islands. These gene set differences between TE islands and LDRs can either be explained by selection processes, removing vital genes from linkage to TE islands or by selective restriction of TE accumulations to genomic regions devoid of such genes.

The current understanding of TE activity dynamics in genomes is that periods of relative dormancy are followed by bursts of activity, often induced by biotic

and abiotic stress, such as exposure to novel habitats. Frequent TE transposition during bursts leads to genomic rearrangements, thus producing new genetic variants and eventually even promoting speciation [66-69]. TE dynamics can also be strongly affected by mating system [3,70-72], and the life history of *C. obscurior* likely challenges the genomic integrity resulting in genomic regions with over 50% TE content. In conclusion, TE dynamics in *C. obscurior* seem to have shifted from a serial to a parallel mode, where a fraction of the genome is reshaped repeatedly in a continuous burst of TE activity. Strikingly, the inbred parasitoid wasp *N. vitripennis* has similar TE frequency patterns suggesting that similar life history strategies and their consequences on Ne and drift can lead to convergent genomic organization. TEs represent a major force in evolution, contributing to the generation of genetic variation especially in species confronted with hurdles like inbreeding or repeated bottlenecks. They furthermore seem to play an important role in the rapid adaption of invasive species to novel environments, making it particularly crucial to understand their origin, function and regulation.

3.5 Methods

3.5.1 Organisms

Live colonies of *C. obscurior* were collected from aborted fruits on coconut trees (*Cocos nucifera*) in Brazil (collected in 2009) and from bark cavities in coral trees (*Erythrina sp.*) in Japan (collected in 2010). The colonies were transferred to Regensburg and placed in plastered petri dishes. Food (honey-soaked shreds of paper; Drosophila or small chunks of Periplaneta americana) and water were provided every 3 days and colonies were kept in incubators under constant conditions (12h 28° light / 12h 24° dark). All animal treatment guidelines applicable to ants under international and German law have been followed. Collecting the colonies that form the basis of the laboratory population used in this study was permitted by the Brazilian Ministry of Science and Technology (RMX 004/02). No other permits were required for this study.

3.5.2 De novo genome assembly

The reference genome is based on one colony that was kept under strict inbreeding in the lab for four generations before extractions. Whole DNA was extracted with CTAB. We extracted DNA from 900 ants, which were pooled to be

sequenced with 454 technology. Extracts of 5, 10 and 30 Brazilian males and 26 Japanese males, respectively, were used for Illumina libraries.

We generated 200 and 500 bp insert libraries with Illumina's TruSeq DNA sample preparation kits from 5 μg of total DNA. Quality control and library preparation were carried out by the KFB sequencing centre of the University Regensburg, sequencing runs were performed by Illumina (Hayward, USA) on a HiSeq2000. Quality control, library preparation and sequencing of 8 and 20 kb long paired end libraries (454, Roche) were carried out by Eurofins MWG Operon (Ebersberg, Germany). Extracted DNA was fragmented into the appropriate fragment sizes (8 and 20 kb) using the HydroShear DNA Shearing Device (GeneMachine). Further library preparation was performed according to 'GS FLX Titanium Paired End Library Prep 20+8 kb Span Method Manual' before sequencing on a GS FLX Titanium (Roche).

The de novo genome assembly was created with MSR-CA version 1.4 open source assembler (University of Maryland genome assembly group). The MSR-CA assembler combines a deBruijn graph strategy with the traditional Overlap-Layout-Consensus employed by various assembly programmes for Sanger-based projects (Arachne, PCAP, CABOG). The MSR-CA uses a modified version of CABOG version 6.1 for contiging and scaffolding. The combined strategy allowed us to natively combine the short 100 bp Illumina reads and longer 454 reads in a single assembly without resorting to an approach that would require one to assemble each type of data separately and then creating a combined assembly.

3.5.3 Mapping

For each lineage, we randomly sampled 140 M 100 bp reads from libraries generated from 26 (JP) and 30 (BR) male pupae. Raw reads were parsed through quality filtration and adapter trimming (Trimmomatic v0.22, options: HEADCROP:7 LEADING:28 TRAILING:28 SLIDINGWINDOW:10:10) and mapped against the BR reference genome with BWA (bio-bwa.sourceforge.net) and Stampy v1.0.21.

3.5.4 Variant calling

SNV calling was carried out combining samtools (samtools.sourceforge.net) and the GATK (www.broadinstitute.org/gatk/) retaining only those variants called consistently by both tools. The final variant set of 553,052 SNVs and 67,987 InDels was stored in a single VCF file. SNVs were annotated with SNPeff

(snpeff.sourceforge.net) to identify non-synonymous and synonymous substitutions.

3.5.5 Calculation of sliding windows

One kb windows of different stats (TEs, exons, SNPs, coverage) were calculated for all scaffolds based on GFF, VCF and SAM files. For GFF and VCF files, custom bash and perl scripts were used to calculated TE and exon bases per 1 kb, and variant calls per 1 kb. Coverage per 1 kb was calculated from SAM files, using samtools' depth algorithm and custom bash and perl scripts. Subsequent processing, calculating of 200 kb sliding windows and plotting of the data was performed with R v3.0.0 (r-project.org).

3.5.6 Gene expression analysis with RNAseq

We extracted whole RNA with the RNeasy Plus Micro kit (Qiagen). Single end Illumina libraries from amplified RNA (Ovation RNAseq system V2) were generated following the manufacturers protocol (Ovation Rapid Multiplexsystem, NuGEN). Sequencing on an Illumina HiSeq1000 at the in-house sequencing centre (KFB, Regensburg, Germany) generated 20M 100 bp reads per sample (Supplementary Table 16). Raw reads were filtered for adapter contamination (cutadapt), parsed through quality filtration (Trimmomatic v0.27, options: LEADING:10 TRAILING:10 SLIDING:4:10 MINLEN:15), and mapped against the reference genome using the tophat2 (v2.0.8) and bowtie2 (v2.1.0) package (–b2-sensitive mode, mapping rate 50%). Gene expression analysis was carried out with DESeq2, based on count tables produced with HTSeq against the Cobs1.4 MAKER annotation (Supplementary Table 16). Genes were considered to be differentially expressed at a false discovery rate < 0.05 and expression values are reported as untransformed base means of read counts per treatment group, after correcting for library size differences ('size factor normalization').

3.6 Bibliography

1. Charlesworth, D., Charlesworth, B. Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. 18, 237–268 (1987).

2. Lynch, M. The Origins of Genome Architecture Sinauer Associates Inc (2007).

3. Charlesworth, D., Wright, S. I. Breeding systems and genome evolution. Curr. Opin. Genet. Dev. 11, 685–690 (2001).

4. Romiguier, J. et al. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J. Evol. Biol. 27, 593–603 (2014).
5. McDonald, J. H., Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652–654 (1991).
6. Lanfear, R., Ho, S. Y. W., Love, D., Bromham, L. Mutation rate is linked to diversification in birds. Proc. Natl Acad. Sci. USA 107, 20423–20428 (2010).
7. Lynch, M. Evolution of the mutation rate. Trends Genet. 26, 345–352 (2010).
8. Fontdevila, A. The Dynamic Genome Oxford Univ. Press (2011).
9. Fedoroff, N. V. Plant Transposons and Genome Dynamics in Evolution John Wiley and Sons (2013).
10. González, J., Karasov, T. L., Messer, P. W., Petrov, D. A. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet. 6, e1000905 (2010).
11. Casacuberta, E., González, J. The impact of transposable elements in environmental adaptation. Mol. Ecol. 22, 1503–1517 (2013).
12. Madlung, A., Comai, L. The effect of stress on genome regulation and structure. Ann. Bot. Lond. 94, 481–495 (2004).
13. Rostant, W. G., Wedell, N., Hosken, D. J. Transposable elements and insecticide resistance. Adv. Genet. 78, 169–201 (2012).
14. Kazazian, H. H. Mobile elements: drivers of genome evolution. Science 303, 1626–1632 (2004).
15. Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., Capy, P. The struggle for life of the genome's selfish architects. Biol. Direct 6, 19 (2011).
16. van Zweden, J. S., D'Ettorre, P. inInsect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology Cambridge Univ. Press (2010).
17. Holt, C., Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491 (2011).
18. Nygaard, S. et al. The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. Genome Res. 21, 1339–1348 (2011).
19. Suen, G. et al. The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 7, e1002007 (2011).
20. Wurm, Y. et al. The genome of the fire ant Solenopsis invicta. Proc. Natl Acad. Sci. USA 108, 5679–5684 (2011).

21. Smith, C. R. et al. Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc. Natl Acad. Sci. USA 108, 5667–5672 (2011).
22. Smith, C. D. et al. Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). Proc. Natl Acad. Sci. USA 108, 5673–5678 (2011).
23. Bonasio, R. et al. Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science 329, 1068–1071 (2010).
24. Weinstock, G. M. et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443, 931–949 (2006).
25. Werren, J. H. et al. Functional and evolutionary Insights from the genomes of three parasitoid Nasonia species. Science 327, 343–348 (2010).
26. Oxley, P. R. et al. The genome of the clonal raider ant Cerapachys biroi. Curr. Biol. 24, 451–458 (2014).
27. Yandell, M., Ence, D. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342 (2012).
28. Gadau, J. et al. The genomic impact of 100 million years of social evolution in seven ant species. Trends Genet. 28, 14–21 (2012).
29. Tsutsui, N. D., Suarez, A. V., Spagna, J. C., Johnston, J. S. The evolution of genome size in ants. BMC Evol. Biol. 8, 64 (2008).
30. Medvedev, P., Stanciu, M., Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods 6, S13–S20 (2009).
31. Treangen, T. J., Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46 (2012).
32. Dubnau, J. et al. The staufen/pumilio pathway is involved in Drosophila long-term memory. Curr. Biol. 13, 286–296 (2003).
33. Millar, N. S., Denholm, I. Nicotinic acetylcholine receptors: targets for commercially important insecticides. Invert. Neurosci. 7, 53–66 (2007).
34. Hemingway, J., Ranson, H. Insecticide resistance in insect vectors of human disease. Annu. Rev. Entomol. 45, 371–391 (2000).
35. Drapeau, M. D., Albert, S., Kucharski, R., Prusko, C., Maleszka, R. Evolution of the yellow/major royal jelly protein family and the emergence of social behavior in honey bees. Genome Res. 16, 1385–1394 (2006).
36. Chen, M.-E., Lewis, D. K., Keeley, L. L., Pietrantonio, P. V. cDNA cloning and transcriptional regulation of the vitellogenin receptor from the imported fire ant, Solenopsis invicta Buren (Hymenoptera: Formicidae). Insect Mol. Biol. 13,

195–204 (2004).

37. Duvernell, D. D., Schmidt, P. S., Eanes, W. F. Clines and adaptive evolution in the methuselah gene region in Drosophila melanogaster. Mol. Ecol. 12, 1277–1285 (2003).

38. Gilbert, L. I. Insect Molecular Biology and Biochemistry Academic Press (2010).

39. Worby, C. A., Dixon, J. E. Sorting out the cellular functions of sorting nexins. Nat. Rev. Mol. Cell Biol. 3, 919–931 (2002).

40. Galiegue, S. et al. Cloning and characterization of PRAX-1. A new protein that specifically interacts with the peripheral benzodiazepine receptor. J. Biol. Chem. 274, 2938–2952 (1999).

41. Kaminker, J. S. et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 3, research0084 (2002).

42. Capy, P., Gasperi, G., Biémont, C., Bazin, C. Stress and transposable elements: co-evolution or useful parasites? Heredity 85, 101–106 (2000).

43. Nuzhdin, S. V., Pasyukova, E. G., Mackay, T. F. Accumulation of transposable elements in laboratory lines of Drosophila melanogaster. Genetica 100, 167–175 (1997).

44. Shee, C., Gibson, J. L., Rosenberg, S. M. Two mechanisms produce mutation hotspots at DNA breaks in Escherichia coli. Cell Rep. 2, 714–721 (2012).

45. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. Science 338, 758–767 (2012).

46. Casida, J. E., Durkin, K. A. Neuroactive insecticides: targets, selectivity, resistance, and secondary effects. Annu. Rev. Entomol. 58, 99–117 (2013).

47. Xu, L., Wu, M., Han, Z. Overexpression of multiple detoxification genes in deltamethrin resistant Laodelphax striatellus (Hemiptera: Delphacidae) in China. PLoS ONE 8, e79443 (2013).

48. Slotkin, T., Seidler, F. Transcriptional profiles reveal similarities and differences in the effects of developmental neurotoxicants on differentiation into neurotransmitter phenotypes in PC12 cells. Brain Res. Bull. 78, 211–225 (2009).

49. Bergé, J. B., Feyereisen, R., Amichot, M. Cytochrome P450 monooxygenases and insecticide resistance in insects. Philos. Trans. R. Soc. Lond. B Biol. Sci. 353, 1701–1705 (1998).

50. Blomquist, G. J., Nelson, D. R., Derenobales, M. Chemistry, biochemistry, and physiology of insect cuticular lipids. Arch. Insect Biochem. Physiol. 6, 227–265 (1987).

51. Foley, B., Chenoweth, S. F., Nuzhdin, S. V., Blows, M. W. Natural ge-

netic variation in cuticular hydrocarbon expression in male and female Drosophila melanogaster. Genetics 175, 1465–1477 (2007).

52. Vosshall, L. B., Wong, A. M., Axel, R. An olfactory sensory map in the fly brain. Cell 102, 147–159 (2000).

53. Zhou, X. et al. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. PLoS Genet. 8, e1002930 (2012).

54. Kulmuni, J., Wurm, Y., Pamilo, P. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. Heredity 110, 538–547 (2013).

55. Guo, S., Kim, J. Molecular evolution of Drosophila odorant receptor genes. Mol. Biol. Evol. 24, 1198–1207 (2007).

56. Hill, C. A. et al. G protein-coupled receptors in Anopheles gambiae. Science 298, 176–178 (2002).

57. Bohbot, J. et al. Molecular characterization of the Aedes aegypti odorant receptor gene family. Insect Mol. Biol. 16, 525–537 (2007).

58. Conceição, I. C., Aguade, M. High incidence of interchromosomal transpositions in the evolutionary history of a subset of or genes in Drosophila. J. Mol. Evol. 66, 325–332 (2008).

59. Tsutsui, N. D., Suarez, A. V., Grosberg, R. K. Genetic diversity, asymmetrical aggression, and recognition in a widespread invasive species. Proc. Natl Acad. Sci. USA 100, 1078–1083 (2003).

60. Feschotte, C. Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 9, 397–405 (2008).

61. Rebollo, R. et al. A snapshot of histone modifications within transposable elements in Drosophila wild type strains. PLoS ONE 7, e44253 (2012).

62. Buchon, N., Vaury, C. RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity 96, 195–202 (2006).

63. Bonasio, R. et al. Genome-wide and caste-specific DNA methylomes of the ants Camponotus floridanus and Harpegnathos saltator. Curr. Biol. 22, 1755–1764 (2012).

64. Hunt, B. G., Glastad, K. M., Yi, S. V., Goodisman, M. A. D. The function of intragenic DNA methylation: insights from insect epigenomes. Integr. Comp. Biol. 53, 319–328 (2013).

65. Veluchamy, A. et al. Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. Nat. Commun. 4, 2091

(2013).

66. Bailey, J. A., Liu, G., Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823–834 (2003).

67. de Boer, J. G., Yazawa, R., Davidson, W. S., Koop, B. F. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics 8, 422 (2007).

68. Ungerer, M. C., Strakosh, S. C., Zhen, Y. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. 16, R872–R873 (2006).

69. Hurst, G. D., Werren, J. H. The role of selfish genetic elements in eukaryotic evolution. Nat. Rev. Genet. 2, 597–606 (2001).

70. Charlesworth, D., Charlesworth, B. Transposable elements in inbreeding and outbreeding populations. Genetics 140, 415–417 (1995).

71. Boutin, T. S., Le Rouzic, A., Capy, P. How does selfing affect the dynamics of selfish transposable elements? Mobile DNA 3, 5 (2012).

72. Wright, S. I., Ness, R. W., Foxe, J. P., Barrett, S. C. H. Genomic consequences of outcrossing and selfing in plants. Int. J. Plant Sci. 169, 105–118 (2008).

73. Wang, J. et al. A Y-like social chromosome causes alternative colony organization in fire ants. Nature 493, 664–668 (2013).

Chapter 4

Genome Analysis of Planctomycetes Inhabiting Blades of the Red Alga *Porphyra umbilicalis*

4.1 Foreword and Acknowledgements

This chapter describes collaborative work that was published in PLoS One in 2016 (Kim JW, Brawley SH, Prochnik S, Chovatia M, Grimwood J, Jenkins J, et al. (2016) Genome Analysis of Planctomycetes Inhabiting Blades of the Red Alga Porphyra umbilicalis. PLoS ONE 11(3): e0151883. https://doi.org/10.1371/journal.pone.0151883). In the following, the main text of the manuscript published in PLoS One is provided with references to Supporting Information, which can be found at the address: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0151883.

I performed nearly all of the data analysis for this work. I wrote the manuscript together with Dr. Arthur Grossman, Dr. Susan Brawley and Dr. John Stiller. Sequencing, assembly and general annotation of the Planctomycete genomes was performed at the Joint Genome Institute by Dr. Simon Prochnik, Dr. Mansi Chovatia, Dr. Jane Grimwood, Dr. Jerry Jenkins, Dr. Kurt LaButti, Dr. Konstantinos Mavromatis, Dr. Matt Nolan, Dr. Matthew Zane and Dr. Jeremy Schmutz.

I would like to give my heartfelt thanks to all of my co-authors, and also to Dr. Nicolas Blouin and Dr. Lilibeth Miranda for their work in maintenance of the

P.um.1 cultures and DNA isolations.

4.2 Introduction

Marine macroalgae and bacteria have varied and complex interactions [1]. Remarkably, the red macroalga *Delisea pulchra* foils attack from a proteobacterium by producing furanones that inhibit quorum-sensing molecules (N-acyl homoserine lactones, AHLs) used for bacterial communication [2]. In contrast, swimming zoospores of the green alga Ulva select settlement sites by sensing AHLs produced by some bacteria [3]. For heterotrophic bacterial "farmers" [4], macroalgal cell walls are a carbon-rich habitat, while bacterial symbionts may synthesize plant growth regulators that stabilize macroalgal morphology [5–7] and provide the algae with inorganic nutrients and vitamins e.g., [8]. Presently, only one symbiotic association has been characterized in some detail i.e. [9].

Little is known as to why different bacteria colonize different algae, and the nature of the complex and dynamic interactions between them [10]. Sympatric macroalgae growing together can harbor substantially different proportions of bacterial phyla. Phyletic effects on the bacterial composition can be larger than observed seasonal or biogeographic impacts [11–13], suggesting that bacteria have selective abilities to feed on different algal cell wall types. Cell wall composition varies among the marine Chlorophyta [cellulose, xyloglucan, mannan, glucuronan, (1,3) β-glucan, ulvan], Rhodophyta [cellulose, (1,4) β-D-mannan, (1,4) β-D-xylan, (1,3) β-D-xylan, glucomannan, sulfated MLG, (1,3) (1,4) β-D-xylan, agars, porphyran, carrageenans] and Phaeophyceae [cellulose, sulfated xylofucoglucuronan, (1,3) β-glucan, alginates (polymannuronic acid, polyguluronic acid), homofucans] [14]. Moreover, within the Rhodophyta, the cell walls in different phases of the life histories (e.g., gametophyte/sporophyte) can show variations in their compositions [14–16].

Examination of the coding capacity of different bacteria for enzymes that degrade cell wall moieties informs our understanding of microbial/macroalgal ecology and evolution. Red algal cell walls are composed mostly of the sulfated polymers porphyran, agar and/or carrageenan, in addition to some xylan and/or cellulose microfibrils [14]. The biosynthesis and degradation of sulfated cell wall polysaccharides of macrophytic algae requires several enzymes including glycoside hydrolases (GHs), sulfatases and carbohydrate sulfotransferases. A range of such enzymes are encoded on many marine bacterial genomes. Indeed, the marine bacterium Zobellia galactanivorans (Bacteroidetes) has a genome encoding 130 GHs, 12 polysaccha-

ride lyases and 71 sulfatases (Genoscope: G0L495), and is being developed as a model for producing enzymes that function in bioconversion of algal polysaccharides. This bacterium is associated with green algae [17,18], red algae [19], brown algae [17] and dinoflagellates [20], and has been examined in detail for its ability to synthesize enzymes capable of degrading sulfated galactans of the red macrophyte Delesseria sanguinea, with specific characterizations of β-agarases [21–23], κ- and ι-carrageenases [24,25] and porphyranases [26,27].

Recently, attention has focused on another macroalgal-associated bacterial phylum, the Planctomycetes. These bacteria are usually a smaller proportion of the macroalgal associated bacteria than Bacteroidetes or Proteobacteria, but may account for 50% of the bacteria on some brown algae e.g., [28]. The Planctomycetes are part of the Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) superphylum [29,30], including some genera that can synthesize a large number of hydrolytic enzymes [31,32]. They exhibit unusual features for bacteria, including division by budding, endocytosis with coated vesicles, a wall composed primarily of glutamine-rich glycoproteins and extensive invaginations of the inner membrane [33–35]. Further, many planctomycete genes are not organized into operons [31], and some encode proteins more typically found in eukaryotes [36].

In a recent study [4], bacterial diversity on the blades of *Porphyra umbilicalis* (Rhodophyta) was analyzed from wild plants and antibiotic-treated, laboratory-cultures. Eight phyla were identified (Bacteroidetes, Proteobacteria, Planctomycetes, Chloroflexi, Actinobacteria, Deinococcus-thermus, Firmicutes, and the candidate division TM7), with the majority of sequences from both field and laboratory material coming from the Bacteroidetes. The abundance of blade-associated Planctomycetes was small on wild blades (0.03–1.1%), but enriched (4.06%,) when *P. umbilicalis* strain P.um.1 [37] was treated with antibiotics that eliminate most bacteria. Four planctomycete OTUs were enriched: *Rhodopirellula baltica* and three undescribed planctomycetes. We have assembled the genomes of these three undescribed planctomycetes and examine their phylogenetic affiliations, genome structures and functional potential.

4.3 Methods

4.3.1 Sample collection

The P.um.1 isolate was collected at Schoodic Point, Maine (40°20'1.68" N; 68°3'29.14"W) on April 3, 2008 [4,37,38]. Details regarding sample preparation are available in S1 Text. Scientific research and collecting permits authorizing field studies pertaining to the P.um.1 isolate were obtained from the United States Department of the Interior, National Park Service, Acadia National Park (permit numbers: ACAD-2008, 2009, 2010, 2011-SCI-0004). These field studies did not involve protected or endangered species.

4.3.2 Genome sequencing and assembly

The 454 sequencing was performed on standard (500–800 bp) and long distance (10 kb) paired-end, genomic libraries (S1 Text). The three largest scaffolds (8.5, 7.3 and 3.8 Mbp) from a preliminary assembly with Newbler (v.2.3-PreRelease-10/20/2009, Roche) were microbial based on sequence similarities in the NCBI (nr) database. We performed additional Illumina sequencing to correct 454 homopolymer errors in the three scaffolds and reassembled the 3.8 Mbp scaffold into a 4.9 Mbp scaffold because it appeared to be an incomplete genome based on its gene complement (S1 Text). These three large scaffolds correspond to genomes of Planctomycetes that we designated P1 (8.5 Mbp), P2 (7.3 Mbp) and P3 (4.9 Mbp).

4.3.3 Genome annotation

The three scaffolds were first annotated through the Joint Genome Institute's microbial annotation pipeline and deposited in the Integrated Microbial Genomes (IMG) database (http://img.jgi.doe.gov/). Additional annotations were conducted for genes of interest with missing functional annotations, protein-coding gene families, repetitive DNA elements, transposable element (TE)-associated genes, selenoproteins and selenocysteine utilization elements, and genomic islands. See S1 Text for additional information.

4.3.4 Phylogenetic analyses

An initial phylogeny based on 16S rDNA sequences for 25 bacterial species was generated using RAxML [39] with the GTR-GAMMA model. A more robust

phylogeny was built by sampling across multiple protein-coding loci [40] corresponding to 39 single-copy genes encoding highly conserved proteins (S1 Table and S1 Text). Homologs for the 39 genes from each of the 23 genomes studied (S2 Table) were aligned, trimmed and then concatenated adhering to a predetermined, randomized gene order. A maximum-likelihood (ML) phylogeny based on 8,725 amino acid positions was inferred from 1000 bootstrap iterations using RAxML. All protein-coding gene trees (see Results) were generated using a similar procedure (S1 Text).

4.3.5 Classification of sulfatases and carbohydrate active enzymes

Sulfatase subclasses were determined based on clades in ML phylogenies of all sulfatase sequences for a given organism. Each resolvable clade was annotated as iduronate-2-sulfatase, heparan-N-sulfatase, mucin-desulfating sulfatase or choline sulfatase, based on BLASTp hits against UNIPROT TREMBL [41]. Unresolvable sulfatases were placed in the more general categories 'arylsulfatase A' and 'galactosamine-N-acetyl-6-sulfatases' (GALNS). We identified hydrolytic enzymes in the Carbohydrate-Active enZYmes (CAZY) database (http://www.cazy.org) using the CAZY Analysis Toolkit, which executes a BLASTp search against the CAZY database. Hits to the genomes used for our analysis had e-values of $< 10^{-10}$.

4.3.6 Identification of genes encoding selenoproteins and Sec insertion and utilization elements

The Sec-insertion and utilization genes (selA, selB, selD, ybbB) were identified by sequence alignments (BLASTp) against known bacterial homologs. Genes potentially encoding selenoproteins were identified on the basis of in-frame opal ('UGA') stop codons, homology searches against known selenoproteins and the presence of SECIS elements. See S1 Text for additional information.

4.4 Results and Discussion

4.4.1 Genome assembly validation and phylogeny

The bacterial strains used here, including the three novel planctomycete genomes recovered from the P.um.1 sequenced libraries are given in S2 Table. Properties of P1 (8.5 Mb), P2 (7.3 Mb) and P3 (4.9 Mb) are provided in Table 1 along with tRNA gene predictions for 29 bacterial genomes, including 22 species from the PVC superphylum, in S3 Table.

For phylogenetic classification, we constructed a high resolution [40] ML tree (Fig 1) based on 39 'core' protein-coding genes (S1 Table). The three sequenced genomes are part of a clade that includes the genera *Blastopirellula*, *Pirellula* and *Rhodopirellula*. P3 is recovered as the most ancestral taxon in this clade, while P2 appears to be an undescribed OTU within the genus *Rhodopirellula*, and P1 shares a direct common ancestor with the *Rhodopirellula* sub-clade. A ML tree based on 16S rDNA (S1 Fig) indicates consistent phylogenetic positions for P1, P2 and P3. P1 and P3 represent new Planctomycetes' genera based on 16S rDNA sequence analysis (S4 Table).

4.4.2 Gene functions and gene family content

The P1, P2 and P3 genomes are non-syntenic with those of other sequenced planctomycete genomes (S2 Fig), and previous work showed that gene content is better preserved than synteny among the Planctomycetes [34]. Many planctomycete genomes have extensive expansions of protein-coding gene families e.g., sulfatases in *Rhodopirellula* [32]; this is also the case for P1, P2 and P3 (S5 Table). Within the Planctomycetes, the percentage of genes belonging to gene families (2 or more) ranged from 36% in *P. mikurensis* to 59% in *S. acidiphila*. Previous studies reported a linear relationship between genome size and percentage of genes in families [42,43]. While most genomes that we analyzed followed this trend, there were several outliers (S3 Fig). Some of the *Rhodopirellula* and P1 have low densities of gene families despite their large genomes, while *K. stuttgartiensis* has high gene family density for a small genome (S2 Text).

Table 4.1: Properties of 9 bacterial genome assemblies including 8 Planctomycetes and 1 marine Bacteroidetes. P1, P2 and P3 were sequenced from a blade of *Porphyra umbilicalis*. Strains of *R. baltica*, *P. mikurensis* and *Z. galactanivorans* (Bacteroidetes) were also present on the blade (based on 16S rDNA analysis). *P. staleyi*, *R. maiorica* and *B. marina* are the closest known relatives of P1, P2 and P3, respectively.

	R. baltica	*R. maiorica*	P2	P1	*P. staleyi*	*B. marina*	P3	*P. mikurensis*	*Z. galactanivorans*
Genome assembly size (kb)	7146	8874	7267	8470	6196	6654	4918	3803	5522
Number of scaffolds	1	1132	1	1	1	64	1	1	1
Estimated size of gaps (kb)	0	0	242	205	0	0	15	0	0
GC content (%)	55.4	54.7	54.9	49.1	57.5	57	61.7	73.3	42.8
Protein-coding genes	7325	7825	5409	6382	4773	6025	4088	3201	4732
rRNA genes	3	3	3	4	3	9	4	3	6
tRNA genes	76	80	51	54	46	53	96	46	40
Other RNA genes	10	**	11	3	3	6	3	2	9
Tandem repeat content (repeat bases / kb of genome)	12.2	18.7	12.7	10.8	16.4	10	19.7	93.5	19.3
TE-associated genes	85	48	26	32	28	75	84	16	28
CRISPRs									
Confirmed	0	0	1	0	2	1	1	2	1
Questionable*	2	11	6	6	0	4	4	13	1

* Small CRISPRs with only two or three direct repeats or CRISPR structures where direct repeats are not 100% identical.

** Data not available in the Integrated Microbial Genomes database.

doi:10.1371/journal.pone.0151883.t001

Highly represented gene families are summarized in S6 Table, with the full list of families in S1 Data. The largest gene families encode response regulators (RR), serine/threonine protein kinases (STPK), transporters (ABC), sigma factors, sulfatases and solute-binding proteins with the 1559 domain of unknown function (DUF1559), which appears exclusive to the PVC superphylum. While some gene families are expanded throughout the Planctomycetes, others such as the sulfatases are more specific to phylogenetic position and/or the type of habitat in which the organism is found (e.g., relative number of sulfatase genes in marine vs. freshwater vs. anammox Planctomycetes).

An investigation of the relationship between higher-level functional classification and gene family size across the 23 genomes studied shows relatively small variations in the COG functional distribution of singleton genes when compared to gene families with more than one member (Fig 2). The largest variation across 23 genomes is in the category 'inorganic ion transport and metabolism' (P), which contains the sulfatases. The absolute distribution of COG domain hits for P1, P2 and P3 is shown in S4 Fig. More in-depth data on gene families and higher-level functional classifications are in the S2 Text.

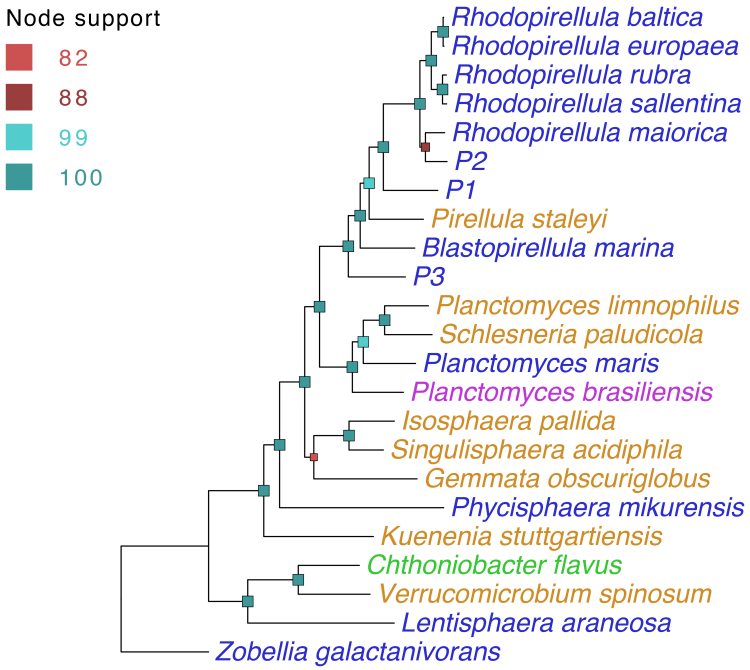

Figure 4.1: Phylogeny of three novel planctomycetes and related species. The phylogeny shown is based on concatenated protein-coding sequences of 39 highly conserved, single-copy genes (see S1 Table). Consensus maximum likelihood trees from 1000 bootstrap iterations are shown. Internal nodes are color-coded (indicated to the left of each tree) based on bootstrap support values. Taxa are color-coded by the type of habitat from which they were isolated: marine [blue], freshwater [orange], marine/brackish [purple], soil [green].

4.4.3 The sulfatases

Sulfatase genes comprise one of the largest families in the Planctomycetes, especially in the genus *Rhodopirellula* (Fig 3, S6 Table). Both sulfatases and GHs are needed for degrading algal cell walls, allowing bacteria to access fixed carbon in sulfated polysaccharides, which can make up in excess of 50% of the dry biomass of macrophytic algae [14,44,45]. Sulfatases catalyze the hydrolysis of sulfate esters and couple with sulfotransferases to facilitate both degradation and synthesis of compounds containing esterified sulfate. The various sulfatases, including alkyl- and arylsulfatases, can have distinct specificities, metabolizing sulfated carbohydrates, proteins and lipids, as well as sulfated glycosaminoglycans and glycolipids [46–48]. A diversity of carbohydrate sulfates can serve as sulfatase substrates, including polysaccharides in cell walls of marine macrophytic algae [27,49,50].

Various sulfatase types are encoded on the planctomycete genomes. Counting only "full-length" ORFs (encoding at least 350 amino acids and containing the active site), there are 122 putative sulfatases in P1, 129 in P2 and only 20 in P3; results for all 23 organisms in our analyses are given in Fig 3a. The active sites of

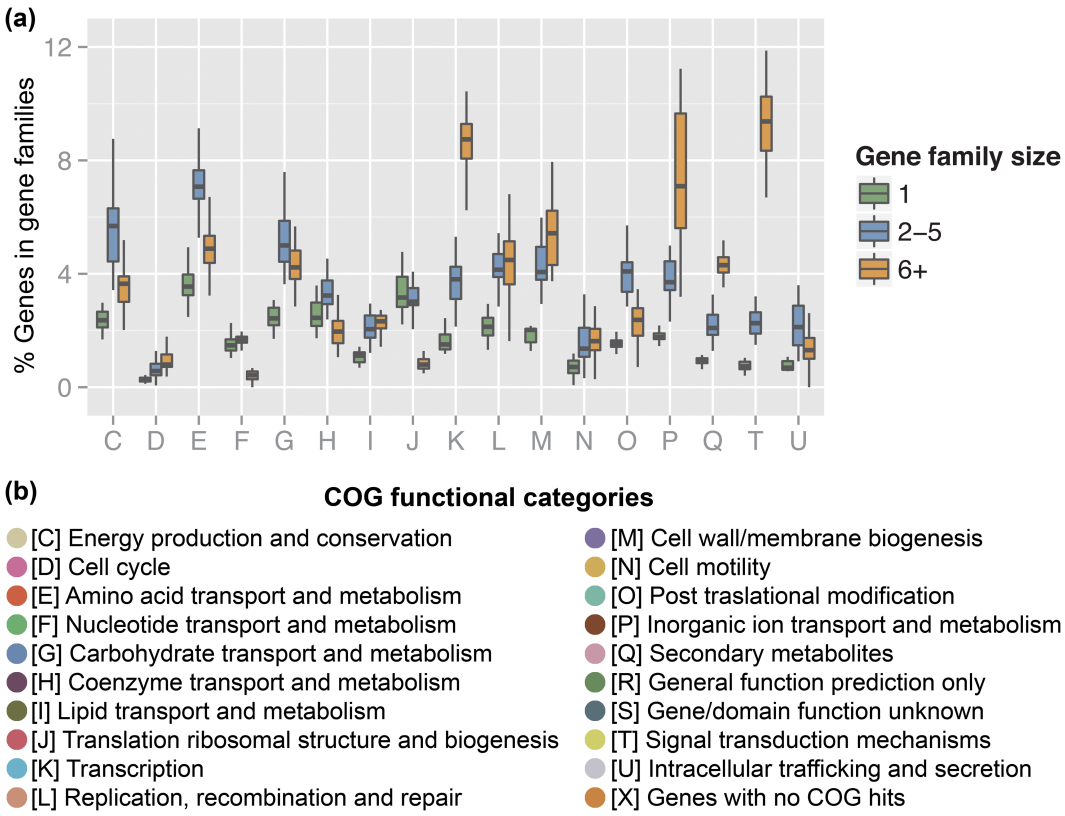

Figure 4.2: Distribution of COG functional categories in paralogous gene families. (a) Distribution across families containing only singletons, or with 2–5 members or 6+ members. Paralogous gene families were identified using a network-based approach (see S1 Text). (b) Definition of COG categories on the x-axis of (a) (and also in S4 Fig).

sulfatases are defined by the sequence C/S-X-P-S/X-R-X-X-X-L/X-T/X-G/X-R/X, in which the cysteine is modified to a formylglycine. The various sulfatases are classified as iduronate-2-sulfatases, heparan-N-sulfatases, mucin-desulfating sulfatases, GALNS sulfatases, with many in the more general arylsulfatase category. The number of full-length sulfatases in each category, determined by phylogenetic analyses, is given in Fig 3b. Based on signal sequence predictions, 79, 91 and 10 sulfatases from P1, P2 and P3, respectively, enter the secretory pathway, likely accessing their substrates from the extracellular space. Enzymes involved in conversion of the sulfatase active site cysteine to a formyl-glycine [51] are also encoded on the P1, P2 and P3 genomes, with 7, 7 and 8 genes, respectively (S2 Text).

While the distribution of sulfatase genes on the P1, P2 and P3 genomes appears to be largely random, some occur in clusters resembling operons (Fig 4). In P1, P2 and P3 there are 10, 20 and 3 instances, respectively, where sulfatase genes reside at adjacent positions on the genome, with a single pair in P1 (IMG: 2643311965, 2643311966) that shows relatively high amino acid sequence identity

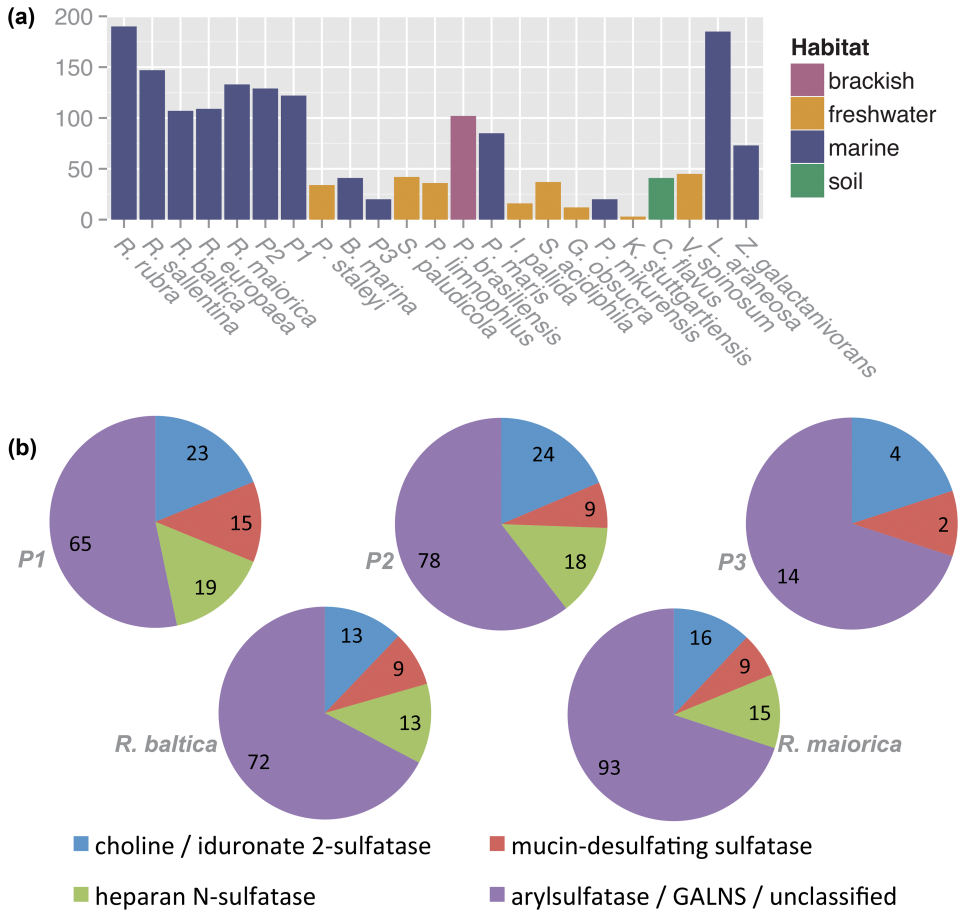

Figure 4.3: Sulfatase gene distribution and sub-classification in Planctomycetes and related strains. (a) Number of sulfatase genes in various Planctomycetes and related strains. Only sulfatase genes encoding the active site and at least 350 amino acid residues were included. (b) Functional subclasses of sulfatases present in P1, P2, P3, *R. baltica* and *R. maiorica*. For each organism, the total number of sulfatases (at least 350 residues) is divided into the following subclasses: choline/iduronate-2-sulfatases, mucin-desulfating sulfatases, heparan-N-sulfatases, and unclassified sulfatases including general arylsulfatases and galactosamine N-acetyl-6-sulfate sulfatases.

(76%) and thus likely arose via a recent tandem duplication. The remaining adjacent pairs are dissimilar (avg. BLASTp sequence identity for P1, 29.0±6.1; P2, 27.6±3.7; P3, 26.7 ± 3.4) and have significantly higher sequence identity to putative PVC orthologs than to each other (avg. BLASTp identity for P1, 64.0 ± 12.7; P2, 68.1 ± 10.6; P3, 51.2 ± 9.9). Also, potential orthologs encoding adjacent P1, P2 and P3 sulfatases are rarely adjacent on the genomes of other closely related Planctomycetes. This suggests that most tandem arrangements of sulfatase genes in P1, P2 and P3 are the consequence of genomic rearrangements and/or HGT, rather than recent tandem duplications. Interestingly, the likelihood of finding even a single pair of adjacent sulfatase genes on the P1, P2 and P3 genomes is very small (permutation test with 10,000 permutations, P1, p = 0.0; P2, p = 0.0; P3, p = 0.0) assuming

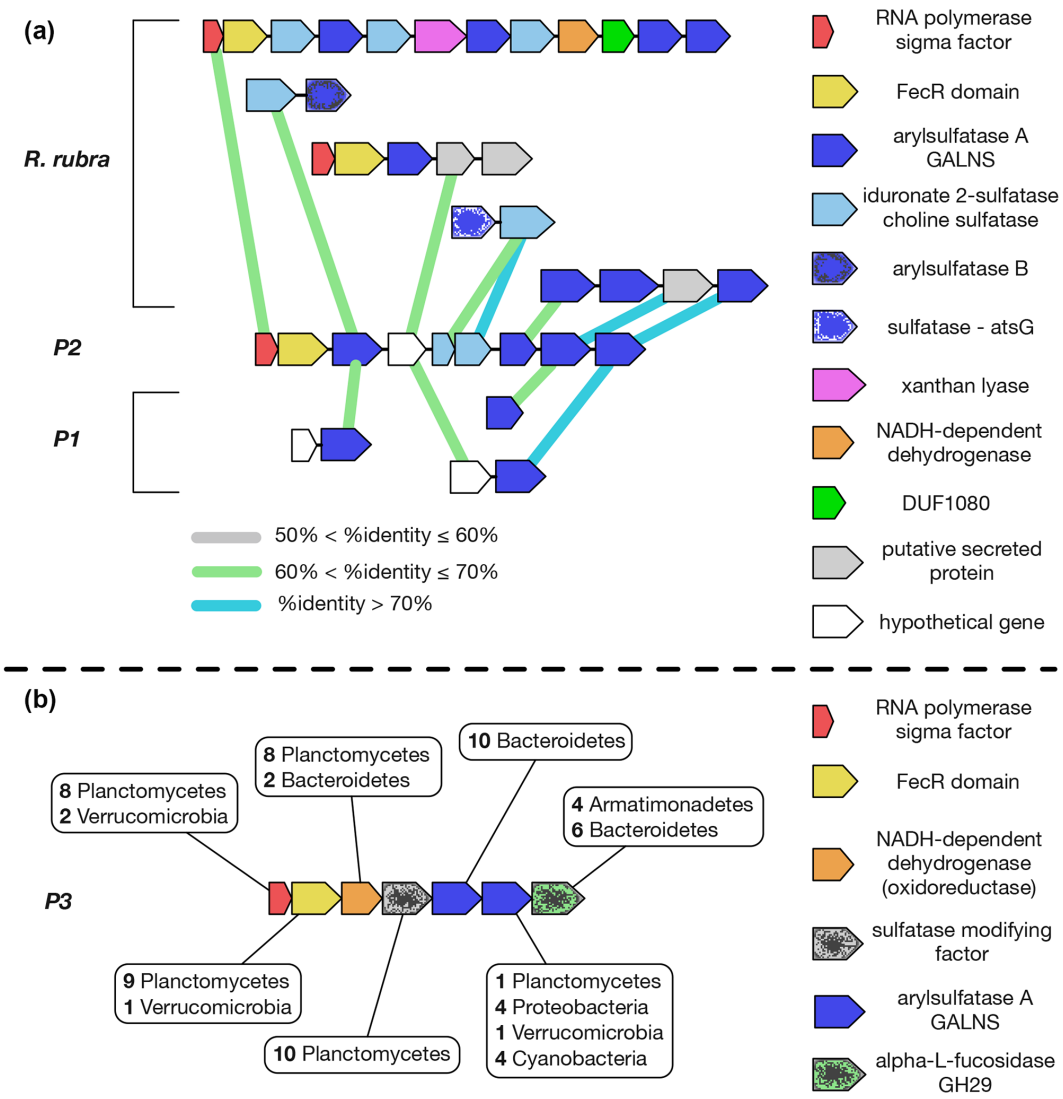

Figure 4.4: Changing context of sulfatase genes in operons. (a) Changing genetic context of individual sulfatase genes of a co-oriented P2 sulfatase gene cluster, resembling an operon. Adjacent genes are joined by a black line, and all genes are color-coded by predicted function as given on the right-hand side of the figure. P1 and *R. rubra* homologs for individual sulfatase genes in the P2 operon are shown. For each homolog, the immediate context of adjacent, co-oriented genes within their respective genomes is also shown. Reciprocal best-hit genes across organisms are connected by thick colored lines (gray, green, cyan). ORF lengths and intergenic distances are not drawn to scale. (b) A heterophyletic gene cluster resembling an operon in P3. Seven consecutive genes are color-coded by predicted function as given on the right-hand side of the figure. The distribution of top 10 BLASTp hits across various bacterial phyla is provided for each gene.

random genome rearrangements with no tandem duplications. This suggests that functional associations (e.g., co-expression of adjacent genes working together to degrade specific polysaccharides) could drive sulfatase gene clustering.

Interestingly, a gene containing two sulfatase domains is present in both P1 (IMG: 2643314295) and P2 (IMG: 2643291516), likely resulting from the fusion

of two unrelated, adjacent sulfatase genes. The two ancestral domains of this gene appear to have different evolutionary origins; the protein encoded by the 5' domain most closely resembles (60% amino acid identity to P1 and P2 orthologs) an arylsulfatase A from the Verrucomicrobia bacterium SCGC AAA164-E04 (GI: 518992481), while the protein encoded by the 3' domain is most similar (50% amino acid identity to P1 and P2 orthologs) to an iduronate-2-sulfatase/choline sulfatase of *Saccharicrinis fermentans* (GI: 763406655) in the Bacteroidetes. We estimate based on protein length (¿600 amino acids) that there are 34 and 21 sulfatase genes on P1 and P2 that encode multi-domain proteins, most often containing glycoside hydrolase and hypothetical protein domains, but also including alginate lyase, esterase/lipase, laminin G, and HEAT_2 repeat domains. Gene fusion appears to contribute to the evolution of multi-domain sulfatase genes, potentially pairing sulfatases with various other functions.

The expansion of the sulfatase gene family appears to be accompanied by high rates of genomic rearrangement consistent with prior observations [34] that can lead to innovation of protein function (e.g., domain swapping and gene fusion) as well as the generation and modification of operons (Fig 4a). In P1, P2 and P3, co-oriented gene clusters resembling operons are often heterophyletic (i.e. member genes with different evolutionary backgrounds). One such P3 gene cluster is shown in Fig 4b, in which member genes, including two sulfatase genes and an α-L-fucosidase gene, have highly discordant BLASTp hit distributions (across NCBI nr); the closest hits for individual members occur in the Bacteroidetes, Proteobacteria, Armatimonadetes and the Planctomycetes. Furthermore, there appears to be a high turnover rate of member genes within such clusters as evidenced by rearrangements of sulfatase genes between various planctomycete OTUs, even within the same genera (Fig 4a). Despite this high turnover rate, likely caused by random genomic rearrangements and HGT, genes encoding polysaccharide degradation enzymes are often found in clusters (e.g., adjacent sulfatase genes, Fig 4a and 4b). One possible explanation is that diversification of operons can confer an adaptive advantage, and is therefore selected.

4.4.4 Polysaccharide degrading enzymes

Sulfated polysaccharides like agars, carrageenans and porphyrans have high proportions of galactose monomers within a polymeric hexose structure. The porphyran polymer, like agarose, has a backbone of repeating disaccharide units, but

the disaccharide is a 3-linked β-D-galactosyl unit alternating with a 4-linked 3,6-anhydro-α-L-galactose. Some of the monomeric units are sulfated at the C6 position while others may be methylated [52]; this is not characteristic of agarose.

Based on P1, P2 and P3 genome sequences, these organisms can synthesize a large number of GHs and polysaccharide lyases (PLs) that have the potential to degrade both 1,3 and 1,4 hexose polymers. GH and PL subclasses that are abundant or over-represented in at least one of the three planctomycete isolates are given in Table 2, with descriptions of the subclasses in S7 Table. Many subclasses are also represented in other Planctomycetes, in members of the larger PVC superphylum, and in *Z. galactanivorans*. The distributions of genes across all CAZY families and subclasses for the 23 genomes are provided in S2 Data.

Enzymes specifically involved in degradation of the Porphyra cell wall include the β-porphyranases in the GH16 subclass and the β-agarases that cleave β-1,4 glycosidic bonds (GH16, GH50, GH86, and GH118) [50]. Genes encoding members of these GH subclasses are unevenly distributed throughout the Planctomycetes. Putative orthologs for GH16 β-porphyranase genes, porA-porE (proteins characterized for *Z. galactanivorans*), are present in some characterized planctomycete genomes, but none encode a full set. *R. rubra* and *R. sallentina* each contain 3 β-porphyranase genes, one of which appears to be porD, while *R. maiorica* has only one ortholog. P3 and *P. mikurensis* each have one β-porphyranase gene, which clade with 72% node support (S5a Fig), while P1 and P2 have no β-porphyranase gene. Genes encoding GH16 β-agarases, such as those of *Z. galactanivorans* (agaA-agaD), are not present in P1, P2 or P3. Within the Planctomycetes, these genes are only in *R. sallentina* (1 gene) and *P. mikurensis* (2 genes); their phylogenetic placement in the context of four *Z. galactanivorans* β-agarase genes is presented in S5a Fig. There are, however, several Planctomycetes with GH50 and GH86 β-agarases, including P1, P2 and P3; GH118 β-agarases are not present in P1, P2 or P3.

GH117 α-neoagarobiases may be keystone enzymes for cleaving α-1,3 glycosidic linkages present in agarose [53]. Proteins of the GH43 subclass, which are structurally related to the GH117s [53], includes galactosidases, xylanases, arabinases and xylosidases, all of which would likely hydrolyze linkages in macroalgal cell walls. Furthermore, GH43 and GH117 proteins appear to be distantly related to the sulfatases based on the high incidence of GH43 and GH117 domain hits (BLASTp e-value $< 10^{-10}$) within sulfatases of P1, P2 and P3.

Table 4.2: Cell wall degradation enzymes in planctomycetes and related species. The number of BLASTp hits (e-value $< 1e^{-10}$) is shown for selected GH and PL domains, which are involved in the degradation of algal, fungal, and vascular plant cell walls. The rows are ordered according to the phylogeny in Fig 1. Entries for P1, P2 and P3 are bolded in cases where the number of members within a CAZY category has a percent rank among all shown species that is greater than 75%.

	cellulases/xylanases			agarases/carrageenases/ galactanases/porphyranases				fucosidases		arabinases/ neoagarobiases		alginate lyases/ pectate lyases		
Organism	**GH3**	**GH10**	**GH74**	**GH16**	**GH50**	**GH53**	**GH86**	**GH29**	**GH95**	**GH43**	**GH117**	**PL6**	**PL9**	**PL14**
Rhodopirellula rubra	7	29	1	17	3	21	1	28	20	**43**	**44**	5	2	0
Rhodopirellula sallentina	4	20	3	15	2	18	4	23	24	50	31	0	2	1
Rhodopirellula baltica	9	17	0	4	0	15	1	1	1	27	12	0	1	1
Rhodopirellula europaea	14	18	0	4	0	15	1	1	1	31	15	0	0	2
Rhodopirellula maiorica	13	19	1	11	1	18	1	11	13	44	29	0	0	0
P2	8	**32**	0	4	1	9	0	2	1	**37**	14	1	**3**	**2**
P1	**25**	**31**	**4**	**18**	**4**	13	**2**	2	1	30	14	**3**	**4**	**4**
Pirellula staleyi	8	8	0	4	0	7	0	0	2	9	3	0	1	0
Blastopirellula marina	12	7	0	3	0	9	0	0	2	12	7	0	1	0
P3	11	4	1	**11**	1	11	1	**8**	**5**	7	4	1	0	0
Schlesneria paludicola	7	5	0	4	0	11	0	0	2	10	4	0	0	0
Planctomyces limnophilus	6	5	0	2	0	6	0	0	0	7	6	0	1	0
Planctomyces brasiliensis	10	13	0	1	1	8	0	1	0	29	14	0	0	0
Planctomyces maris	8	9	0	1	1	8	0	0	2	20	7	1	3	0
Isosphaera pallida	6	1	1	4	0	5	0	1	2	4	2	0	2	0
Singulisphaera acidiphila	21	5	1	3	0	11	0	0	3	12	2	0	0	0
Gemmata obscuriglobus	11	3	0	6	0	0	0	0	0	5	2	0	1	0
Phycisphaera mikurensis	4	8	2	6	2	3	4	1	1	6	2	0	1	0
Kuenenia stuttgartiensis	3	0	0	0	0	24	0	0	0	0	0	0	0	0
Chthoniobacter flavus	23	5	3	8	1	0	0	0	1	14	6	0	2	0
Verrucomicrobium spinosum	13	12	0	5	2	0	0	0	3	11	7	1	0	1
Lentisphaera araneosa	23	35	0	7	3	0	1	10	13	74	50	0	1	0
Zobellia galactanivorans	22	8	2	19	0	0	0	20	8	15	16	2	2	1

doi:10.1371/journal.pone.0151883.t002

The GH16 subclass includes genes encoding κ-carrageenases, which are found in P1 (IMG: 2643316630), *L. araneosa*, *Z. galactanivorans*, and the *Rhodopirellula*, including P2 (IMG: 2643292705). Genes putatively encoding ι-carrageenases are present only in *R. rubra* while λ-carrageenases are found in *R. rubra* and *R. sallentina*. While carrageenan is not present in *Porphyra umbilicalis* or any other member of the Bangiophyceae, it is the main cell wall polysaccharide of the red alga *Chondrus crispus*. In most areas of the North Atlantic, including Maine where P.um.1 was collected (S6 Fig), *P. umbilicalis* is positioned only 1–2 vertical meters from rich Chondrus beds.

Several of the investigated genomes also contain multiple genes encoding enzymes that potentially degrade fucans and alginates in brown algal cell walls. For instance, the GH29 (α-1,3/1,4-L-fucosidase) and GH95 (α-1,2-L-fucosidase) subclasses are highly expanded in 3 out of 6 members of *Rhodopirellula*, while the other three genomes, including P2, only contain 1 or 2 genes for these proteins (Table

2). The GH29 and GH95 subclasses have also expanded in P3, *L. araneosa*, and *Z. galactanivorans*. P1 and P2 contain multiple genes encoding PL6, PL9 and PL14 alginate lyases, while P3 has only a single gene member in PL6.

Some GH subclasses are represented by either zero or low membership in P1, P2 and P3. For example, P2 has no members in GH74, GH86 and GH118. There can also be major differences in the number of members of specific GH subclasses in the Planctomycetes [e.g., from 74 to 0 for GH43 and from 50 to 0 for GH117 (Table 2)]. Furthermore, 13 GH subclasses have maximum and minimum representations across the 6 *Rhodopirellula* genomes that differ by 10 or more members.

Cell wall polysaccharides comprise the majority of dry biomass of marine macroalgae, providing a rich carbon source for heterotrophic bacteria. Within meters of each other in the rocky intertidal and shallow subtidal zones of the North Atlantic shore are red algae with cell walls rich in carrageenan or agar rather than porphyran, brown algal kelps (subtidal) and rockweeds (high to low intertidal) that contain sulfated fucans and alginate, green macroalgae that have ulvans (sulfated glucuronoxylorhamnogalactans) and, especially in brown and green macroalgae, considerable cellulose [16]. It is unclear how much specificity there is in the cell-wall digesting capability of macroalgal-associated bacteria, but genomic analyses of their wall digesting capabilities may help explain their relative abundances on different groups of marine algae. Furthermore, substrate availability also impacts expression of the bacterial hydrolytic genes. When grown on the brown algal carbohydrate reserve laminarin, *Z. galactanivorans* expresses *porA* and *porB*, which encode enzymes that cleave neoporphyranobiose (L6S-G) in agar polymers [50]. However, when *Z. galactanivorans* is grown on a red alga with an agar-containing wall, the *agaA*, *agaB*, *agaC* and *agaD* genes are expressed, while a porphyran substrate elicits expression of *agaA*, *agaB*, *agaC*, *porC* and *porE* [50].

Variation in distribution of different GH categories among the three different planctomycete isolates raises the possibility that these bacteria have preferred niches [30] among the macroalgae. For example, P3 appears to be adapted to degrading brown algal cell walls based on the large number of fucosidases encoded in its genome; these have low representation in P1 and P2 (Table 2). P1 and P2 both appear well-equipped to live on both green and red algal cell walls based on their expanded arsenal of cellulases, arabinases, xylanases, agarases, porphyranases, galactanases, and carrageenases; GH10 xylanases comprise one of the largest expansions in P1 and P2 (Table 2).

4.4.5 Horizontal gene transfer

Expansion of protein-coding gene families involving intra-chromosomal gene duplications (IGD) and horizontal gene transfers (HGT) is a key component of adaptive evolution. The relative impacts of IGD and HGT on bacterial evolution have been debated [42,54], with likely different roles in niche adaptation for paralogs acquired through IGD and xenologs acquired through HGT [55].

In general, definitive evidence for HGT is difficult to obtain; however, support can be acquired through various semi-quantitative metrics involving comparisons against "true" evolutionary lineages (as predicted in Fig 1). These metrics include (1) high bootstrap support for heterophyletic clades except in cases of long-branch attraction [56], and (2) markedly higher sequence identity to gene(s) in more distantly related organisms than to orthologs in close relatives. Using such metrics, we predict numerous instances of HGT between the Planctomycetes and other bacterial/archaeal phyla and also between different genera within the Planctomycetes. Here we highlight cases of potential HGT in P1, P2 and P3 that appear to be associated with niche adaptations.

HGT of genes encoding polysaccharide-degrading enzymes can reflect adaptation to colonizing specific macroalgae. For instance, P1 appears to have acquired its ability to degrade κ-carrageenan from the Bacteroidetes; the P1 κ-carrageenase protein (IMG: 2643316630) clades with *Z. galactanivorans* and *C. drobachiensis* (98% node support) (S5b Fig), and is more similar in amino acid sequence to the protein of *Z. galactanivorans* (63% identity over 95% length) than to the closest planctomycete hit [*R. europaea* (GI: 460274492) at 44% identity]. Also, the phylogeny of eight α-L-fucosidases (GH29) in P3 is indicative of mixed evolutionary origins (Fig 5a). Only one of the eight fucosidases is terminally cladded to another planctomycete (*R. sallentina*), while the others have their closest known relatives in Bacteroidetes, Armatimonadetes, and Gemmatimonadetes. Finally, both P1 and P2 show expansions in the family of PL14 alginate polysaccharide lyases, where a pair of P1 and P2 genes exhibits high amino acid sequence identity (74%), indicating a strong possibility for HGT of these genes (Fig 5b). HGT from free-living marine Bacteroidetes is known to have played a significant role in increasing degradative capability of marine Proteobacteria for digesting alginates [50] and for introducing genes encoding enzymes involved in alginate and porphyran digestion into human gut Bacteroidetes [26,50].

Genes in the planctomycete genomes potentially involved in adaptation to

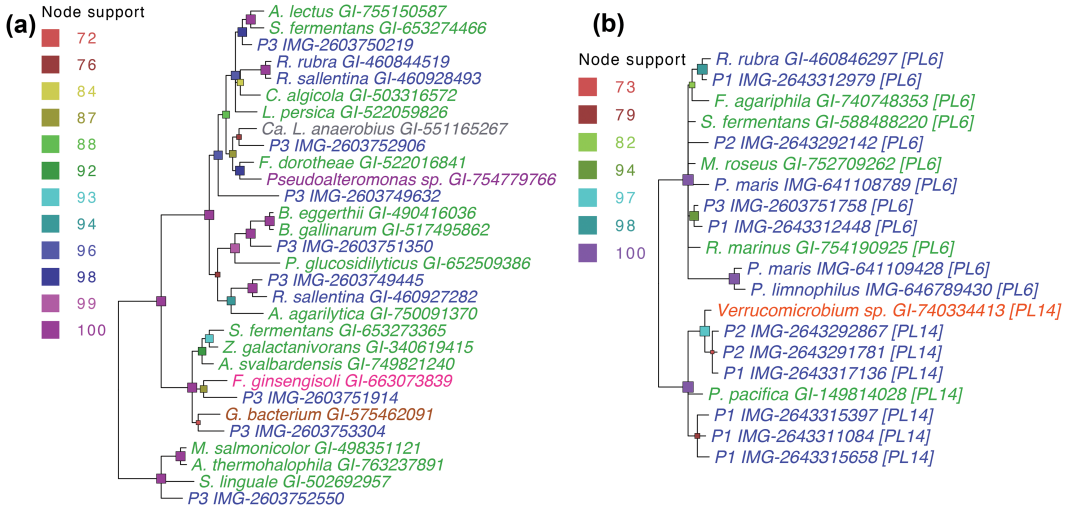

Figure 4.5: Phylogenies of polysaccharide-degrading enzymes indicate host adaptation. (a) Phylogeny of P3 α-L fucosidases (GH29). The genes included in the phylogeny are top hits having more than 50% sequence identity at 80% query coverage that were determined by BLASTp of each of the eight P3 fucosidases to the NCBI nr database and to the genomes included in this study. (b) Phylogeny of PL6 and PL14 alginate lyases. The genes included in the phylogeny are top hits having more than 50% sequence identity at 80% query coverage that was determined by the BLASTp of each of the P1, P2 and P3 alginate lyases to the NCBI nr database and to the genomes included in this study. In both (a) and (b), genes are color-coded by organism as follows: Planctomycetes [blue], Bacteroidetes [green], Proteobacteria [purple], Verrucomicrobia [red-orange], Armatimonadetes [magenta], Gemmatimonadetes [brown], unclassified [gray]. Node support is from 1000 bootstrap iterations.

environmental stress are those most likely acquired by HGT. Multi-drug efflux pumps (pfam00873) are responsible for ejecting environmental and intracellular toxins such as metabolites, dyes, detergents, bile salts and antibiotics from cells. In *E. coli*, mutations in genes associated with TolC-dependent efflux systems cause up-regulation of various stress responses in *E. coli* [57]. P1 and P2 both contain a gene for an AcrB-type efflux pump, which is an inner membrane component of a TolC system. These P1 and P2 genes (IMG: P1–2643312425, P2–2643289582) encode proteins that are highly similar in sequence (83% identity over 99% of length) and do not appear to have vertically transmitted homologs in other Planctomycetes including *Rhodopirellula*, the genus to which P2 belongs (S7a Fig). The next closest match to the P2 protein is encoded by *R. maiorica*, at 52% sequence identity. These observations could reflect recent HGT between P1 and P2, or sequence convergence driven by purifying selection from shared environmental pressures reflecting variation in substrate specificities.

Amino acid transporters can be part of cellular stress response mechanisms, including those of the acid resistance system in *E. coli* [58], salt-stress induction of

proline transporters in yeast [59], and the eukaryotic response to protein synthesis inhibition by oxidative stress [60]. A highly conserved amino acid transporter in P1 and P2 (66% amino acid identity over 99% of the length; IMG: P1–2643312291, P2–2643289856), but not encoded on any of the other planctomycete genomes, displays homology to transporters encoded on the genomes of a few members of the Bacteroidetes and Proteobacteria, and more broadly to various halophilic archaeal genomes (S7b Fig); these findings suggest the occurrence of HGT from Archaea to Bacteria, and then among a few bacterial phyla including the Planctomycetes. While the physiological role of this transporter is not known, it may function in response to frequent stresses in the intertidal zone, including high salinity and the absorption of excess excitation energy.

Genomic islands (GI) are horizontally transmitted gene clusters, generally mediated by transposable elements (TEs), that can facilitate adaptation to specific environments by conferring a selective advantage to the recipient [61]. P1, P2 and P3 contain putative GIs that span 4.2, 187.1 and 248.7 kbp, respectively. P3 has the largest number of TE-associated genes (Table 1) and also contains the largest total GI region (S8 Fig, S1 and S2 Texts). Functional predictions and the distributions of P1, P2 and P3 genes occurring in GIs are available in S3 Data. Notably, one of the P3 GH29 α-L-fucosidases (IMG: 2603749632) occurs in a GI. In addition, P1, P2 and P3 and many other Planctomycetes contain degenerate tRNA gene clusters with large numbers of partially degraded tRNAs, which are often acquired through HGT and thus, may be dispensable to the carrier organism [62,63]. Perhaps the most notable horizontal acquisition by the Planctomycetes is of a highly canonical isoleucine tRNA gene (tRNA-UAU) that occurs as a single-copy within degenerate tRNA gene clusters in several planctomycete genomes, including P1, P2 and P3. Codon usage analysis suggests that tRNA-UAU facilitates the translation of more recently acquired genes (such as genes in GIs), thereby increasing the rate at which new protein functions are established (S2 Text).

4.4.6 Selenoproteins in P1 and P2

Adaptation to stress conditions has also been associated with selenoproteins, or enzymes containing selenocysteine (Sec) amino acid residues that generally confer increased catalytic efficiency compared to their sulfur-based, cysteine-containing homologs [64–66]. Most known selenoproteins have redox functions [67], and it has been suggested that the increased catalytic activities of selenoproteins

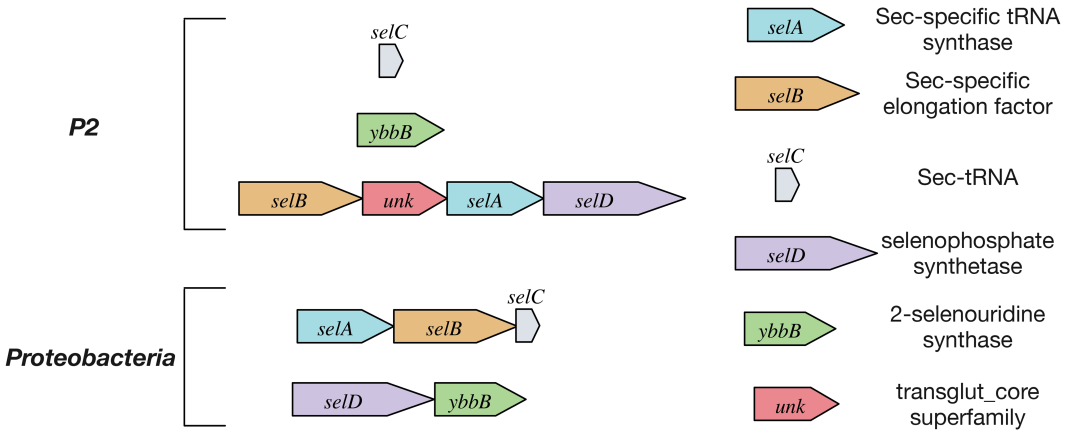

Figure 4.6: Comparison of operons encoding genes required for selenocysteine insertion and selenophosphate synthesis/utilization. For Proteobacteria, the two operons shown generally represent conserved structures for the majority of Sec-encoding and Sec-utilizing proteobacterial species. The Sec-insertion operon structure shown for P2 has not been found in other known genomes (NCBI), including P1. An additional gene is shown that contains a transglut_core domain (PFAM001841; likely to have cysteine protease function in prokaryotes).

are most beneficial in extreme environments associated with high levels of oxidative stress [68]. The largest known selenoproteome belongs to the harmful pelagophyte *Aureococcus anophagefferens* [69]. This picoplankton occurs in dense estuarine blooms where a portion of the cells are exposed to high light, elevated temperatures and osmotic stress [70]. Exposure to excessive light causes algae to produce reactive oxygen species, which must be quickly detoxified to avoid cellular damage [71].

Selenocysteines are co-translationally inserted into proteins by the selenosome complex [72], which requires 4 dedicated selenocysteine-associated genes (S8 Table). P1 and P2 both contain full sets of genes required for Sec-insertion during protein synthesis as well as genes for 2-selenouridine synthase (S8 Table), which improves base-pair discrimination in select tRNAs. P2 has an operon-like arrangement of these genes that is unusual in comparison to Sec-insertion operons in Proteobacteria (Fig 6), the phylum with the most known selenoproteomes (S2 Text); Sec-insertion genes in P1 are not co-localized. Also, Sec-insertion genes of both P1 and P2 appear to have mixed evolutionary origins (S2 Text). Two other planctomycetes, *G. obscuriglobus* and *I. pallida*, contain full sets of genes required for Sec-insertion, but neither of these genomes contain genes for 2-selenouridine synthase (S8 Table). We did not find genetic evidence for selenocysteine usage in P3.

Genes encoding putative selenoproteins in P1 and P2 were identified as described in the Methods, and are listed in S9 Table. In P1, a formate dehydrogenase α subunit (*fdhA*) is one of six putative selenoproteins with antioxidant activity. In

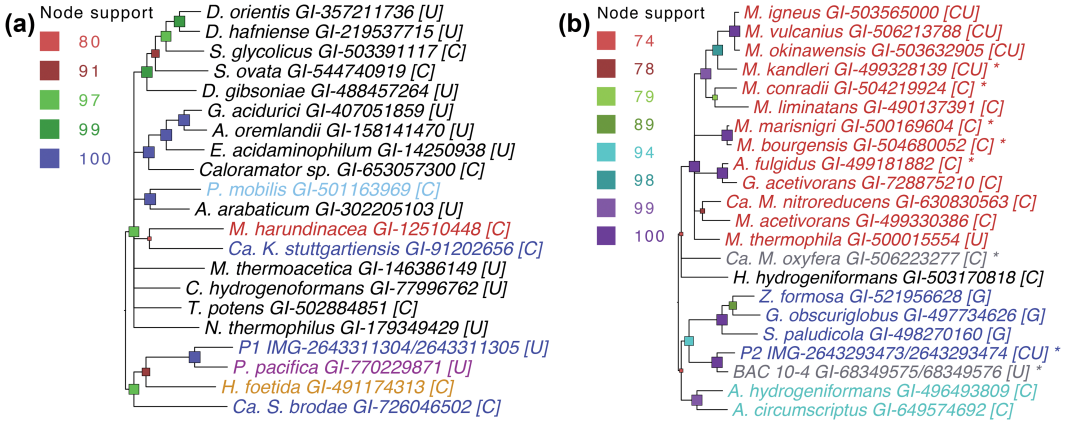

Figure 4.7: Horizontal gene transfer of selenoprotein genes reflects adaptation to stress conditions. (a) Phylogeny of formate dehydrogenase α subunit (*fdhA*). Closest non-redundant hits (BLASTp against NCBI nr) to the P1 selenoprotein sequence are shown. (b) Phylogeny of formylmethanofuran dehydrogenase β subunit (*fmdB*). Closest non-redundant hits to the P2 selenoprotein sequence are shown. Asterisk indicates a similar formylmethanofuran dehydrogenase operon structure as in P2 (*fmdD-fmdB-fmdA-fmdC*). In both (a) and (b), genes are color-coded by organism as follows: Planctomycetes [blue], Proteobacteria [purple], Acidobacteria [orange], Thermotogae [light blue], Firmicutes [black], and Archaea [red], Synergistetes [cyan], unclassified [gray]. Sequences containing selenocysteine are marked with [U] and cysteine-containing sequences are marked with [C]. Node support is from 1000 bootstrap iterations.

Proteobacteria, *fdhA* is generally located near the Sec-insertion operon and may play a role in maintaining the Sec-insertion and decoding traits in bacteria [73]. In P1, the Sec-insertion genes and *fdhA* are not co-localized, but instead, *fdhA* forms an operon with *nuoEF*, genes that encode NADH:ubiquinone dehydrogenase I chains E and F (not selenoproteins). This P1 *fdhA* operon is well conserved (65% amino acid identity) in the myxobacterium *Plesiocystis pacifica* SIR-1 (a proteobacterium isolated from beach seagrass, *Zostera sp.*), but not in any other genome (in NCBI). Phylogenetic analysis indicates that the *fdhA* gene was part of multiple HGT events involving the Planctomycetes, including HGT between the P1 and *P. pacifica* lineages (Fig 7a).

Fig 7b shows a phylogeny of formylmethanofuran dehydrogenase β subunit (*fmdB* gene), which is encoded as a selenoprotein in P2, several Archaea, and two unclassified bacteria; glycine-containing homologs occur in three other Planctomycetes: *G. obscuriglobus*, *S. paludicola*, and *Z. formosa*. The closest match (71% amino acid identity) to P2 *fmdB* is on a fosmid associated with an uncultured bacterium from the freshwater lake, Lake Washington [74]. In P2 and the Lake Washington bacterium, *fmdB* is part of the *fmdD-fmdB-fmdA-fmdC* operon; this operon structure also occurs in several Archaea as well as Candidatus *Methylomirabilis oxyfera*. Some

organisms, including P2, contain both selenocysteine and cysteine-forms of *fmdB*. In *M. kandleri*, these two forms are differentially expressed in response to selenium availability [75].

4.5 Conclusion

This work has revealed numerous metabolic adaptations to the life style of planctomycete colonists of macroalgae within the intertidal zone, including the presence of large families of genes encoding sulfatases and hydrolases that degrade polysaccharides, multidrug transporters, and selenoproteins. Many of the hydrolytic enzymes allow P1, P2 and P3 to feed on the cell walls of the three major macroalgal groups (brown, green and red algae), but there are also suggestions of specialization for specific macroalgal hosts. Evidence for extensive HGT from the Bacteroidetes and Proteobacteria to the Planctomycetes emphasizes the intimate associations among these groups of bacteria on the macroalgal thallus. The interactions of the bacteria with each other, and with their associated macroalgae, are likely to reflect important physiological interactions that allow for the successful cohabitation of the bacteria and alga, and also offer the potential for genetic exchange that continually tailors bacteria to changing environmental conditions and macroalgal distributions.

4.6 References

1. Goecke F, Labes A, Wiese J, Imhoff JF (2010) Chemical interactions between marine macroalgae and bacteria. Marine Ecology Progress Series 409: 267–299.

2. Fernandes N, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2012) Community structure and functional gene profile of bacteria on healthy and diseased thalli of the red seaweed Delisea pulchra. PLoS One 7: e50854. pmid:23226544

3. Tait K, Joint I, Daykin M, Milton DL, Williams P, Camara M (2005) Disruption of quorum sensing in seawater abolishes attraction of zoospores of the green alga Ulva to bacterial biofilms. Environ Microbiol 7: 229–240. pmid:15658990

4. Miranda LN, Hutchison K, Grossman AR, Brawley SH (2013) Diversity and abundance of the bacterial community of the red Macroalga Porphyra umbilicalis: did bacterial farmers produce macroalgae? PLoS One 8: e58269. pmid:23526971

5. Spoerner M, Wichard T, Bachhuber T, Stratmann J, Oertel W (2012) Growth and thallus morphogenesis of Ulva mutabilis (Chlorophyta) depends on a combination

of two bacterial species excreting regulatory factors. J Phycol 48: 1433–1447.

6. Provasoli L, Carlucci AF (1974) Vitamins and growth regulators. Botanical Monographs: 741–787.

7. Fries L (1975) Some observations on the morphology of Enteromorpha linza (L.) J. Ag. and Enteromorpha compressa (L.) Grev. in axenic culture. Bot Mar 18: 251–253.

8. Kazamia E, Czesnick H, Nguyen TT, Croft MT, Sherwood E, Sasso S, et al. (2012) Mutualistic interactions between vitamin B12 -dependent algae and heterotrophic bacteria exhibit regulation. Environ Microbiol 14: 1466–1476. pmid:22463064

9. Matsuo Y, Imagawa H, Nishizawa M, Shizuri Y (2005) Isolation of an algal morphogenesis inducer from a marine bacterium. Science 307: 1598. pmid:15761147

10. Lage OM, Bondoso J (2014) Planctomycetes and macroalgae, a striking association. Front Microbiol 5: 267. pmid:24917860

11. Burke C, Thomas T, Lewis M, Steinberg P, Kjelleberg S (2011) Composition, uniqueness and variability of the epiphytic bacterial community of the green alga Ulva australis. ISME J 5: 590–600. pmid:21048801

12. Lachnit T, Meske D, Wahl M, Harder T, Schmitz R (2011) Epibacterial community patterns on marine macroalgae are host-specific but temporally variable. Environ Microbiol 13: 655–665. pmid:21078035

13. Longford SR, Tujula NA, Crocetti GR, Holmes AJ, Holmström C, Kjelleberg S, et al. (2007) Comparisons of diversity of bacterial communities associated with three sessile marine eukaryotes. Aquatic Microb Ecol 48: 217–229.

14. Popper ZA, Michel G, Herve C, Domozych DS, Willats WG, Tuohy MG, et al. (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. Annu Rev Plant Biol 62: 567–590. pmid:21351878

15. Mukai LS, Craigie JS, Brown RG (1981) Chemical composition and structure of the cell walls of the conchocelis and thallus phases of Porphyra tenera (Rhodophyceae) J Phycol 17: 192–198.

16. Kloareg B, Quatrano RS (1988) Structure of the cell walls of marine algae and ecophysiological functions of the matrix polysaccharides. Oceanogr Mar Biol 26: 259–315.

17. Nedashkovskaya OI, Suzuki M, Vancanneyt M, Cleenwerck I, Lysenko AM, Mikhailov VV, et al. (2004) Zobellia amurskyensis sp. nov., Zobellia laminariae sp. nov. and Zobellia russellii sp. nov., novel marine bacteria of the family Flavobacteriaceae. Int J Syst Evol Microbiol 54: 1643–1648. pmid:15388723

18. Matsuo Y, Suzuki M, Kasai H, Shizuri Y, Harayama S (2003) Isolation and

phylogenetic characterization of bacteria capable of inducing differentiation in the green alga Monostroma oxyspermum. Environ Microbiol 5: 25–35. pmid:12542710
19. Barbeyron G, L'Haridon S, Corre E, Kloareg B, Potin P (2001) Zobellia galactanovorans gen. nov., sp. nov., a marine species of Flavobacteriaceae isolated from a red alga, and classification of [Cytophaga] uliginosa (ZoBell and Upham 1944) Reichenbach 1989 as Zobellia uliginosa gen. nov., comb. nov. Int J Syst Evol Microbiol 51: 985–997. pmid:11411725
20. Skerratt JH, Bowman JP, Hallegraeff G, James S, Nichols PD (2002) Algicidal bacteria associated with blooms of a toxic dinoflagellate in a temperate Australian estuary. Mar Ecol Prog Ser 244: 1–15.
21. Allouch J, Jam M, Helbert W, Barbeyron T, Kloareg B, Henrissay B et al. (2003) The three-dimensional structures of two beta-agarases. J Biol Chem 278: 47171–47180. pmid:12970344
22. Allouch J, Helbert W, Henrissat B, Czjzek M (2004) Parallel substrate binding sites in a beta-agarase suggest a novel mode of action on double-helical agarose. Structure 12: 623–632. pmid:15062085
23. Jam M, Flament D, Allouch J, Potin P, Thion L, Kloareg B, et al. (2005) The endo-beta-agarases AgaA and AgaB from the marine bacterium Zobellia galactanivorans: two paralogue enzymes with different molecular organizations and catalytic behaviours. Biochem J 385: 703–713. pmid:15456406
24. Barbeyron T, Gerard A, Potin P, Henrissat B, Kloareg B (1998) The kappa-carrageenase of the marine bacterium Cytophaga drobachiensis. Structural and phylogenetic relationships within family-16 glycoside hydrolases. Mol Biol Evol 15: 528–537. pmid:9580981
25. Barbeyron T, Michel G, Potin P, Henrissat B, Kloareg B (2000) iota-Carrageenases constitute a novel family of glycoside hydrolases, unrelated to that of kappa-carrageenases. J Biol Chem 275: 35499–35505. pmid:10934194
26. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 464: 908–912. pmid:20376150
27. Correc G, Hehemann J-H, Czjzek M, Helbert W (2011) Structural analysis of the degradation products of porphyran digested by Zobellia galactanivorans β-porphyranase A. Carbohydr Polym 83: 277–283.
28. Bengtsson MM, Øvreås L (2010) Planctomycetes dominate biofilms on surfaces of the kelp Laminaria hyperborea. BMC Microbiol 10: 261. pmid:20950420
29. Hou S, Makarova KS, Saw JH, Senin P, Ly BV, Zhou H, et al. (2008) Com-

plete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. Biol Direct 3: 26. pmid:18593465

30. Bondoso J, Balague V, Gasol JM, Lage OM (2014) Community composition of the Planctomycetes associated with different macroalgae. FEMS Microbiol Ecol 88: 445–456. pmid:24266389

31. Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, et al. (2003) Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. Proc Natl Acad Sci U S A 100: 8298–8303. pmid:12835416

32. Wegner CE, Richter-Heitmann T, Klindworth A, Klockow C, Richter M, Achstetter T, et al. (2013) Expression of sulfatases in Rhodopirellula baltica and the diversity of sulfatases in the genus Rhodopirellula. Mar Genomics 9: 51–61. pmid:23273849

33. Santarella-Mellwig R, Franke J, Jaedicke A, Gorjanacz M, Bauer U, Budd A, et al. (2010) The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol 8: e1000281. pmid:20087413

34. Jogler C, Waldmann J, Huang X, Jogler M, Glockner FO, Mascher T, et al. (2012) Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in Planctomycetes by comparative genomics. J Bacteriol 194: 6419–6430. pmid:23002222

35. Santarella-Mellwig R, Pruggnaller S, Roos N, Mattaj IW, Devos DP (2013) Three-dimensional reconstruction of bacteria with a complex endomembrane system. PLoS Biol 11: e1001565. pmid:23700385

36. Jenkins C, Kedar V, Fuerst JA (2002) Gene discovery within the planctomycete division of the domain Bacteria using sequence tags from genomic DNA libraries. Genome Biol 3: research0031.0031—research0031.0011.

37. Blouin NA, Brawley SH (2012) An AFLP analysis of clonality in widespread asexual populations of Porphyra umbilicalis (Rhodophyta) with a sensitivity analysis for bacterial contamination. Mar Biol 159: 2723–2729.

38. Blouin NA (2010) Asexual reproduction in Porphyra umilicalis Kutzing and its development for use in mariculture: University of Maine. 151 p. 39. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463. pmid:15608047

40. Hofstetter V, Miadlikowska J, Kauff F, Lutzoni F (2007) Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case

study of the Lecanoromycetes (Ascomycota). Mol Phylogenet Evol 44: 412–426. pmid:17207641

41. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011: bar009.

42. Pushker R, Mira A, Rodriguez-Valera F (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol 5: R27. pmid:15059260

43. Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H, et al. (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4: e5299. pmid:19390573

44. Michel G, Tonon T, Scornet D, Cock JM, Kloareg B (2010) The cell wall polysaccharide metabolism of the brown alga Ectocarpus siliculosus. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. New Phytol 188: 82–97. pmid:20618907

45. Kraan S (2012) Algal Polysaccharides, Novel Applications and Outlook. In: Chang CF, editor. Carbohydrates—Comprehensive Studies on Glycobiology and Glycotechnology.

46. Hanson SR, Best MD, Wong CH (2004) Sulfatases: structure, mechanism, biological activity, inhibition, and synthetic utility. Angew Chem Int Ed Engl 43: 5736–5763. pmid:15493058

47. Ghosh D (2007) Human sulfatases: a structural perspective to catalysis. Cell Mol Life Sci 64: 2013–2022. pmid:17558559

48. Toesch M, Schober M, Faber K (2014) Microbial alkyl- and aryl-sulfatases: mechanism, occurrence, screening and stereoselectivities. Appl Microbiol Biotechnol 98: 1485–1496. pmid:24352732

49. Michel G, Nyval-Collen P, Barbeyron T, Czjzek M, Helbert W (2006) Bioconversion of red seaweed galactans: a focus on bacterial agarases and carrageenases. Appl Microbiol Biotechnol 71: 23–33. pmid:16550377

50. Hehemann JH, Correc G, Thomas F, Bernard T, Barbeyron T, Jam M, et al. (2012) Biochemical and structural characterization of the complex agarolytic enzyme system from the marine bacterium Zobellia galactanivorans. J Biol Chem 287: 30571–30584. pmid:22778272

51. Dierks T, Schmidt B, von Figura K (1997) Conversion of cysteine to formylglycine: a protein modification in the endoplasmic reticulum. Proc Natl Acad Sci U S A 94: 11963–11968. pmid:9342345

52. Zhang Q, Qi H, Zhao T, Deslandes E, Ismaeli NM, Molloy F, et al. (2005) Chemical characteristics of a polysaccharide from Porphyra capensis (Rhodophyta).

Carbohydr Res 340: 2447–2450. pmid:16150429
53. Hehemann JH, Smyth L, Yadav A, Vocadlo DJ, Boraston AB (2012) Analysis of keystone enzyme in Agar hydrolysis provides insight into the degradation (of a polysaccharide from) red seaweeds. J Biol Chem 287: 13985–13995. pmid:22393053
54. Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7: e1001284. pmid:21298028
55. Alm E, Huang K, Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol 2: e143. pmid:17083272
56. Stiller JW, Hall BD (1999) Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol 16: 1270–1279. pmid:10939894
57. Rosner JL, Martin RG (2013) Reduction of cellular stress by TolC-dependent efflux pumps in Escherichia coli indicated by BaeSR and CpxARP activation of spy in efflux mutants. J Bacteriol 195: 1042–1050. pmid:23264577
58. Harding HP, Zhang Y, Zeng H, Novoa I, Lu PD, Calfon M, et al. (2003) An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. Mol Cell 11: 619–633. pmid:12667446
59. Rentsch D, Hirner B, Schmelzer E, Frommer WB (1996) Salt stress-induced proline transporters and salt stress-repressed broad specificity amino acid permeases identified by suppression of a yeast amino acid permease-targeting mutant. Plant Cell 8: 1437–1446. pmid:8776904
60. Lu P, Ma D, Chen Y, Guo Y, Chen GQ, Deng H, et al. (2013) L-glutamine provides acid resistance for Escherichia coli through enzymatic release of ammonia. Cell Res 23: 635–644. pmid:23337585
61. Preston GM, Haubold B, Rainey PB (1998) Bacterial genomics and adaptation to life on plants: implications for the evolution of pathogenicity and symbiosis. Curr Opin Microbiol 1: 589–597. pmid:10066526
62. Puerto-Galan L, Vioque A (2012) Expression and processing of an unusual tRNA gene cluster in the cyanobacterium Anabaena sp PCC 7120. Fems Microbiol Lett 337: 10–17. pmid:22924345
63. Levican G, Katz A, Valdes J, Quatrini R, Holmes DS, Orellana O (2009) A 300 kpb genome segment, including a complete set of tRNA genes, is dispensable for Acidithiobacillus ferrooxidans. Biohydrometallurgy: A Meeting Point between Microbial Ecology, Metal Recovery Processes and Environmental Remediation 71–73: 187–190.

64. Arner ES (2010) Selenoproteins-What unique properties can arise with selenocysteine in place of cysteine? Exp Cell Res 316: 1296–1303. pmid:20206159
65. Gromer S, Johansson L, Bauer H, Arscott LD, Rauch S, Ballou DP, et al. (2003) Active sites of thioredoxin reductases: why selenoproteins? Proc Natl Acad Sci U S A 100: 12618–12623. pmid:14569031
66. Johansson L, Gafvelin G, Arner ES (2005) Selenocysteine in proteins-properties and biotechnological use. Biochim Biophys Acta 1726: 1–13. pmid:15967579
67. Zhang Y, Fomenko DE, Gladyshev VN (2005) The microbial selenoproteome of the Sargasso Sea. Genome Biol 6: R37 pmid:15833124
68. Arbogast S, Ferreiro A (2010) Selenoproteins and protection against oxidative stress: selenoprotein N as a novel player at the crossroads of redox signaling and calcium homeostasis. Antioxid Redox Signal 12: 893–904. pmid:19769461
69. Gobler CJ, Lobanov AV, Tang YZ, Turanov AA, Zhang Y, Doblin M, et al. (2013) The central role of selenium in the biochemistry and ecology of the harmful pelagophyte, Aureococcus anophagefferens. ISME J 7: 1333–1343. pmid:23466703
70. Sieburth JM, Johnson PW, Hargraves PE (1988) Ultrastructure and ecology of Aureococcus anophagefferens gen. et sp. nov. (Chrysophyceae): the dominant picoplankter during a bloom in Narragansett bay, Rhode Island, Summer 1985. J Phycol 24: 416–425.
71. Waring J, Klenell M, Bechtold U, Underwood GJC, Baker NR (2010) Light-induced responses of oxygen photoreduction, reactive oxygen species production and scavenging in two diatom species. J Phycol 46: 1206–1217.
72. Hatfield DL, Gladyshev VN (2002) How selenium has altered our understanding of the genetic code. Mol Cell Biol 22: 3565–3576. pmid:11997494
73. Romero H, Zhang Y, Gladyshev VN, Salinas G (2005) Evolution of selenium utilization traits. Genome Biol 6: R66. pmid:16086848
74. Vorholt JA, Kalyuzhnaya MG, Hagemeier CH, Lidstrom ME, Chistoserdova L (2005) MtdC, a novel class of methylene tetrahydromethanopterin dehydrogenases. J Bacteriol 187: 6069–6074. pmid:16109948
75. Vorholt JA, Vaupel M, Thauer RK (1997) A selenium-dependent and a selenium-independent formylmethanofuran dehydrogenase and their transcriptional regulation in the hyperthermophilic Methanopyrus kandleri. Mol Microbiol 23: 1033–1042. pmid:9076739