University of Groningen

Bachelor Thesis

The Adaptive Function of Tycheposons in Marine Bacteria

Author: Diana Jecu *(s3736180)*
Cohort: 2018 Ecology & Evolution

Supervisors:
Dr. Thomas Hackl
Prof. Dr. Joana Falcao Salles

June 26, 2021

rijksuniversiteit
groningen

1 Summary

Mobile genetic elements (MGEs) are a type of genetic material which move within and between genomes and usually carry adaptive functions. They are transmitted between bacteria through horizontal gene transfer (HGT), which is a primary mechanism of bacterial evolution. Tycheposons, which were first found in Prochlorococcus, are elements with site-specific integrases adapted to a mobile life-style. They are split into two categories, by function: PICI-like and cargo-carrying. This study looks into the ecological functionality of the genes of cargo-carrying tycheposons. Tycheposons were searched for in two data sets: Tara Oceans, with 2595 genomic sequences, and GORG, with 12715 genomic sequences. Both are diverse collections of metagenomic data of marine microorganisms from all over the world. In the first set, tycheposons were found in 5.12% of the searched genomes, while in the second set they were only found in 0.32%. The main annotations of the cargo genes in these tycheposons are limiting nutrient uptake (nitrogen, iron, and phosphorus), heavy metal binding (zinc, copper), temperature shock management, and oxidative stress resistance. All of these functions facilitate the adaptation of the host to harsh environmental conditions, be it nutrient limitation or external stressors such as UV radiation, heat, or anthropogenic factors as pollution and potentially even climate change. However, due to the limitations of the sequencing method used for both data sets, further research is needed using other data sets to better examine tycheposon cargo functionality in marine bacteria.

Contents

1 Summary **1**

2 Introduction **3**

3 Methods **7**

4 Results **10**

4.1 Tara Oceans 10

4.2 GORG 12

5 Discussion and Conclusions **14**

6 Appendices **22**

6.1 Appendix 1 - Tara Oceans 22

6.2 Appendix 2 - GORG 32

2 Introduction

Horizontal gene transfer (HGT) is a mechanism through which microbes exchange genetic data, promoting adaptation to different environmental conditions and the diversification of populations. This process is fundamental for microbial evolution [1], and it often involves mobile genetic elements (MGEs). Thanks to HGT, bacteria can evolve rapidly through the acquisition of new and advantageous traits [2]. The canonical mechanisms of HGT are the conjugation of plasmids, phage transduction mediated by bacteriophages, and transformation by naked DNA [3]. Other mechanisms have also been described, both involving MGEs and independent of them. The first include strategies involving elements that mobilize only their own DNA and those that mobilize bacterial DNA, while the latter involve membrane vesicles, autolysis, and bacterial mating [4].

Mobile genetic elements are fragments of genetic material with the ability to move within and between genomes through proteins which they encode [5]. This mobility allows them to alter protein coding regions in the genome of their host and change their function. They often carry beneficial genes for functions which help the host more easily adapt to its environment [6]. In this way, they are important vectors of adaptation and evolution in bacteria. One of the best known examples of mobile genetic element relevance is that of the antibiotic resistance genes transported from one bacterium to neighbouring ones through plasmids - which are circular extrachromosomal DNA molecules [7]. Besides plasmids, some examples of mobile genetic elements include transposons, introns, and phage-inducible chromosomal islands, also known as PICIs. PICIs are a type of mobile genetic element with the ability manipulate phage life cycles in order to promote their own spread [8], and are an important reference element very similar in function to one of the tycheposon categories to be described in the next paragraphs.

This study will look into the functions of a specific type of mobile genetic elements - the newly described tycheposons [9]. They were discovered while searching for mobile genetic element activity in the genomic islands present in the genomes of several Prochlorococcus populations. Horizontal gene transfer often gives rise to genomic islands, which are large chromosomal regions highly variable from one strain to another [9]. These regions are usually diverse assemblages of genetic material previously horizontally transferred, including MGEs, or the remainders thereof in the case where the integration is evolutionarily older and their functionality has been lost [9]. These regions are relevant to explore when looking into the genetic variability of related strains, because this variability largely occurs within them [10]. Tycheposons were discovered during the investigation of Prochlorococcus genomic islands, which are well-established even though there were no known mobile genetic elements or signs of typical horizontal gene transfer in these microorganisms.

It was found that Prochlorococcus islands contain two distinct gene pools characterized by dif-

ferent rates of turn-over and transfer. The first gene pool, which is also the smaller one, contains highly mobile element genes, including specific hallmark genes for integration and replication, along with cargo genes. The second and larger gene pool encompasses island genes which serve no particular function, hypothesized to be flanking material temporarily captured onto an excising element through an inexact excision from the genome of origin [9]. Following the search, few already known elements were observed. However, 937 putative integrase-carrying transposons were found which had not been previously described in the literature. These elements, named "tycheposons" after the Greek goddess Tyche, the patron of oceans, were found mostly adjacent to seven island-associated tRNA genes: Proline (TGG), Serine (TGA), Alanin (GGC), Threonine (GGT), Arginine (TCT), one of three Methionines (CAT), and the tmRNA gene [9].

Two functional types of tycheposons have been observed: cargo-carrying and PICI-like ones. The elements observed so far in the first category predominantly play a role in the capacity of assimilation of limiting nutrients from the environment, promoting niche specialization in the host organisms. The ones in the second category contain viral packaging genes likely used in hijacking phage capsids to be used for the tycheposon's further dispersal. PICIs, or phage-inducible chromosomal islands, represent a class of mobile genetic elements which carry and disperse genes for virulence and antibiotic resistance [11]. They excise from bacterial chromosomes, replicate, and are then encased in particles resembling phages, leading to frequent transfers, both intra- and inter-genomic [12]. Through their functions of encoding regulatory genes for phage infection detection and genes for phage replication interference, they increase the resistance to infections of the host population. As such, PICI-like tycheposons most likely also have a similar beneficial effect on their hosts. Examples of the two types of tycheposons can be seen in Figure 1.

Tycheposons are elements with functions in defense and adaptation. What makes them different from other MGEs, apart from the enzyme used to catalyze their movement, is that they encode site-specific integrases, and many carry genes used for independent replication, such as primases, polymerases, and helicases. Furthermore, they share a conserved gene structure with an inward-facing integrase at one end, optionally a small replication module neighboring the integrase or at the other end, and they carry cargo useful for adaptation to the local environment [9]. While these are not generally uncommon characteristics in MGEs, they are more commonly seen in eukaryotic elements [13] or in larger prokaryotic ones, rather than in small prokaryotes such as Prochlorococcus [9].

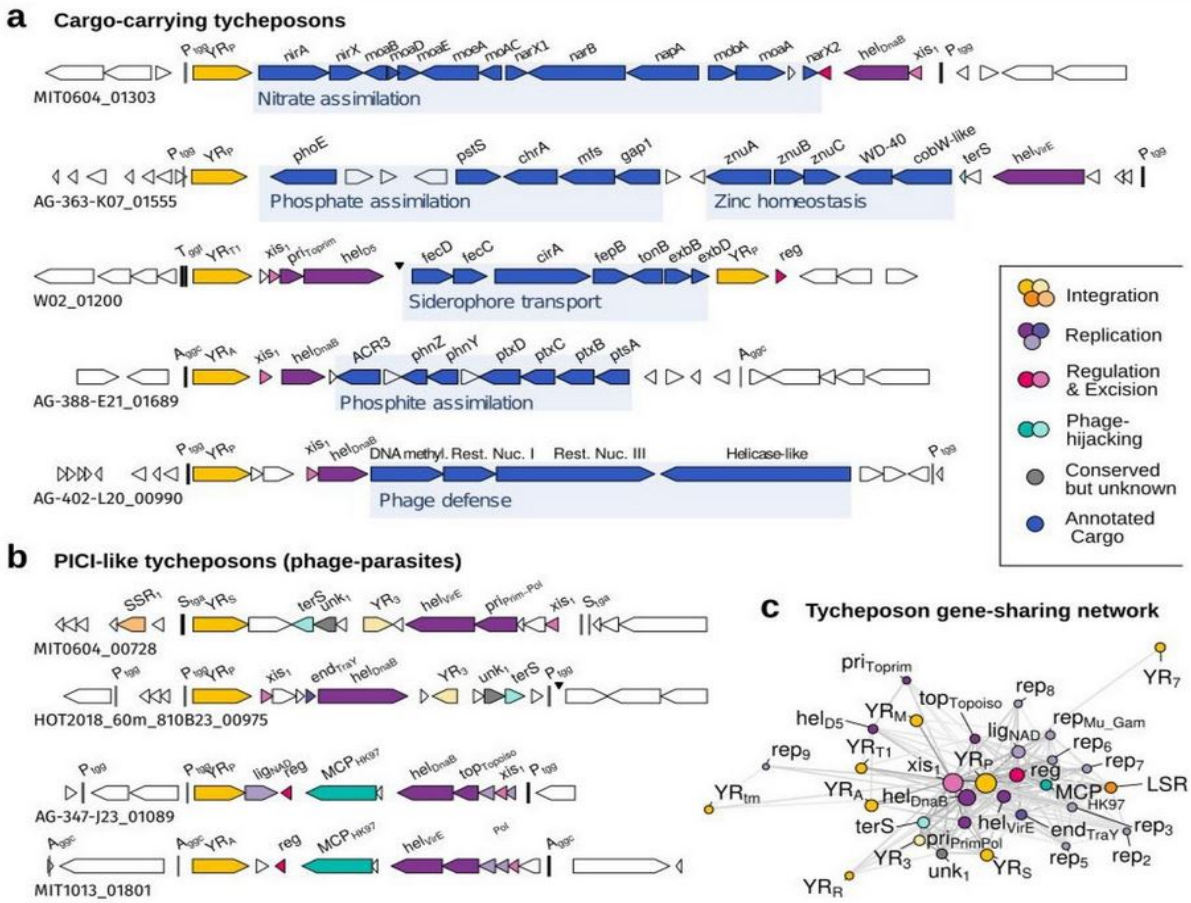


Figure 1: The modular structure of the two categories of tycheposons: a. Five examples of cargo-carrying tycheposons and their functions; b. Four examples of PICI-like tycheposons; c. Tycheposon gene-sharing network showcasing the evolutionary link between the different types, indicated by their integration and replication modules. From Hackl *et al.*, 2020 [9].

The previous research project explored two data sets: Cyanorak, containing other Cyanobacteria aside from Prochlorococcus, and Tara Oceans, containing marine bacteria collected by an expedition with the same name. The project focused on determining the prevalence and the functionality of tycheposons in the genomes contained in these two previously unexplored data sets. To find tycheposons, the genomes were searched against the tycheposon protein sequences provided in the original study [9], and the genes in the matching loci were further annotated to allow for categorization. It was revealed that these MGEs can be found outside of the Prochlorococcus genus, and even outside of the Cyanobacteria phylum. When loooking at the distribution of the two functional categories, in general, cargo-carrying tycheposons are more prevalent than PICI-like ones, possibly due to the more direct advantages they offer to their host. Aside from that, tycheposons within Cyanobacteria appear to be phylogenetically similar, suggesting that horizontal gene transfer of tycheposons can and does easily occur within this phylum. However, the annotations for the cargo carried by the cargo-carrying tycheposons were not added during

the previous study, so the type of genes they carry is still unknown.

As a result of time limitations, the findings of this previous study leave many topics unexplored, and give rise to plenty of new questions. Among these questions, a notable one is "What are the ecological implications of the functions of the cargo carried by tycheposons?", which is what this study aims to explore. To expand the search for tycheposons to more aquatic microorganisms, a new data set - GORG - was explored in a similar fashion as in the previous research project. The annotations of the cargo-carrying tycheposons which were missing from the first study will be added to the already existing Tara Oceans results, as well as to the results of the new search.

From what has been found in the original study, the genes of the cargo-carrying type of tycheposons seem to have limiting nutrient uptake as their main function. Based on these previous findings, it is to be expected that the cargo annotated in this study will be of this type. The main nutrients which limit growth in the ocean are nitrogen, phosphorus, silica, iron, and other metals such as copper [14]. These nutrients limit the primary production in oceanic environments, and they are replenished either by the process of nutrient recycling or through mixing with deeper waters. The replenishment of these nutrients is mostly restricted to decomposition from the atmosphere and deposit inputs from the land along the coast. These elements are all fundamentally vital for aquatic life. Phosphorous and nitrogen are important elements of cells, with the first being a component of cell membrane structures and of ATP, and also a key factor in processes such as gene transcription regulation [15], while the latter is an element of nuclear acids [16]. Furthermore, iron is necessary for a lot of enzymes and processes, for example in photosynthesis [17]. Thus, tycheposons carrying this type of cargo would bring a tremendous advantage to their host genomes in nutrient scarce oceanic environments.

The findings of this study may prove useful for deepening our understanding of how tycheposons have influenced the ecology of bacteria in oceanic conditions, and could also provide insights into the specifics of microbial niche differentiation in marine environments. As they are newly discovered, a lot of information that would help determine their significance remains undiscovered. Still, since their primary functions allow for better adaptation of their host to the environment, they have the potential to be used in predicting the fitness and adaptation capabilities of the host species, especially in nutrient-scarce environments.

3 Methods

The previously explored data set contains 2595 genomes sequenced from data gathered by the Tara Oceans expedition from oceans all around the world, which were downloaded from the European Nucleotide Archive website [18]. The cargo-carrying tycheposon loci discovered in this data set were used in this study. The new data set of GORG contains 12715 partial genomes from the tropical and subtropical euphotic ocean, obtained through single-cell sequencing [19]. This method of sequencing is relatively new and consists of capturing a single cell and amplifying the genetic material contained within to obtain a genomic sequence [20]. Since the GORG data set is significantly larger and more varied than the Tara Oceans one, but its sequences are also less complete, the prevalence of tycheposons in this case is expected to be lower. The reference tycheposon protein sequences of the hallmark genes and their hmmer profiles were taken from the original paper describing the new MGEs [9] available on Github [21]. The protein sequences of the hallmark genes were computed from the protein profiles.

Because of the computational cost of the necessary methods, to optimize efficiency, the data processing required for this study was done via the Peregrine High Performance Computing (HPC) cluster in the Linux command line. Some of the technical requirements of this search included installing the Linux (Ubuntu distribution) operating system, learning how to use the Bash scripting language, and learning the Peregrine HPC methodology. This included learning the appropriate syntax for the commands, the use of sbatch to submit jobs to the cluster, and the use of slurm files for checking the status of the HPC jobs and for checking for possible processing errors.

The workflow overview can be seen in Figure 2. The initial search for tycheposons from the previous project was repeated again for the GORG data set. The search required the use of several programs. Of these, the most used one was mmseqs, a software which can search and cluster large sets of sequences. This program can reach the same sensitivity as BLAST, but it is up to 400 times faster [22], making it a more suitable alternative for larger data sets. In the first necessary step for the search, mmseqs was used to convert the set of tycheposon genes into a database. This database was then clustered into groups of similar sequences through a cascaded clustering algorithm. Profiles of the hallmark gene clusters were then computed and added to a profile database using mmseqs. The ensuing search for tycheposon genes in the genomes was done using these profiles.

The resulting files were imported into R Studio, along with the databases containing the already known tycheposon hallmark genes and the annotations associated therewith, taken from the initial study. In the beginning, the data was filtered to eliminate hits with an e-value lower than 10^{-10}. To be able to create plots to visualize the data, it was necessary to associate the

discovered genes with their respective annotations. However, as the two tables containing the required data have different sequence IDs due to them having been computed using two different programs, a hmmer ID needed to be added to the mmseqs search result table. To add a hmmer ID to the IDs used by mmseqs, a table constructed by matching the hmmer table from the original paper, which also contains the sequence IDs, to the tycheposon gene database was merged to the mmseq search result table by their keys. The annotations available from the initial study were then added to the resulting table, to differentiate between the two tycheposon categories.

The package gggenomes was used to visualize the data and create figures of the hits of genes found. First of all, a locus was defined as a region in the genomic sequence made up of tycheposon gene agglomerations of at least two genes with different roles. These genes are hits found through the mmseqs search, and their position coordinates (start and end locations) in the result table makes it possible to place them in the schematic representation of the genome and to find the closely positioned genes to cluster into a locus. The deciding factor when defining a locus is the presence of at least one tycheposon-specific integrase. These gene agglomerations are the tycheposon-like loci to be further investigated. The function "locate" of the gggenomes package is used in order to find these loci in the genomic sequences based on the criteria described above, and then the function "focus" is used to zoom in on each locus and add its gene structure.

Further on, for both of the data sets, the results are split into two functional categories: PICI-like and non-PICI-like, based on the presence or absence of the packaging modules required for the production of capsids. In the plots, the genes required to indicate tycheposon presence were colored differentially, with their color indicating their function. The genomes with no clear tycheposon-like formations were filtered out and were not included in the final results.

In the next phase of the research, the non-PICI-like loci were cropped out of their sequences using the seqkit software [23], to allow for an accurate and easy addition of the missing cargo annotations. As this study focuses on the functions of the cargo-carrying tycheposons, only this category of loci was further annotated. These annotations come from the Pfam data base, which was downloaded using mmseqs. The tycheposon loci sequences were matched against Pfam and the resulting table was imported into R. Given the large number of different annotations, the tycheposons were not visualised through plotting. A table was constructed for each data set, containing the ID of the original locus sequence, the Pfam ID of the discovered proteins, and the description of their functions.

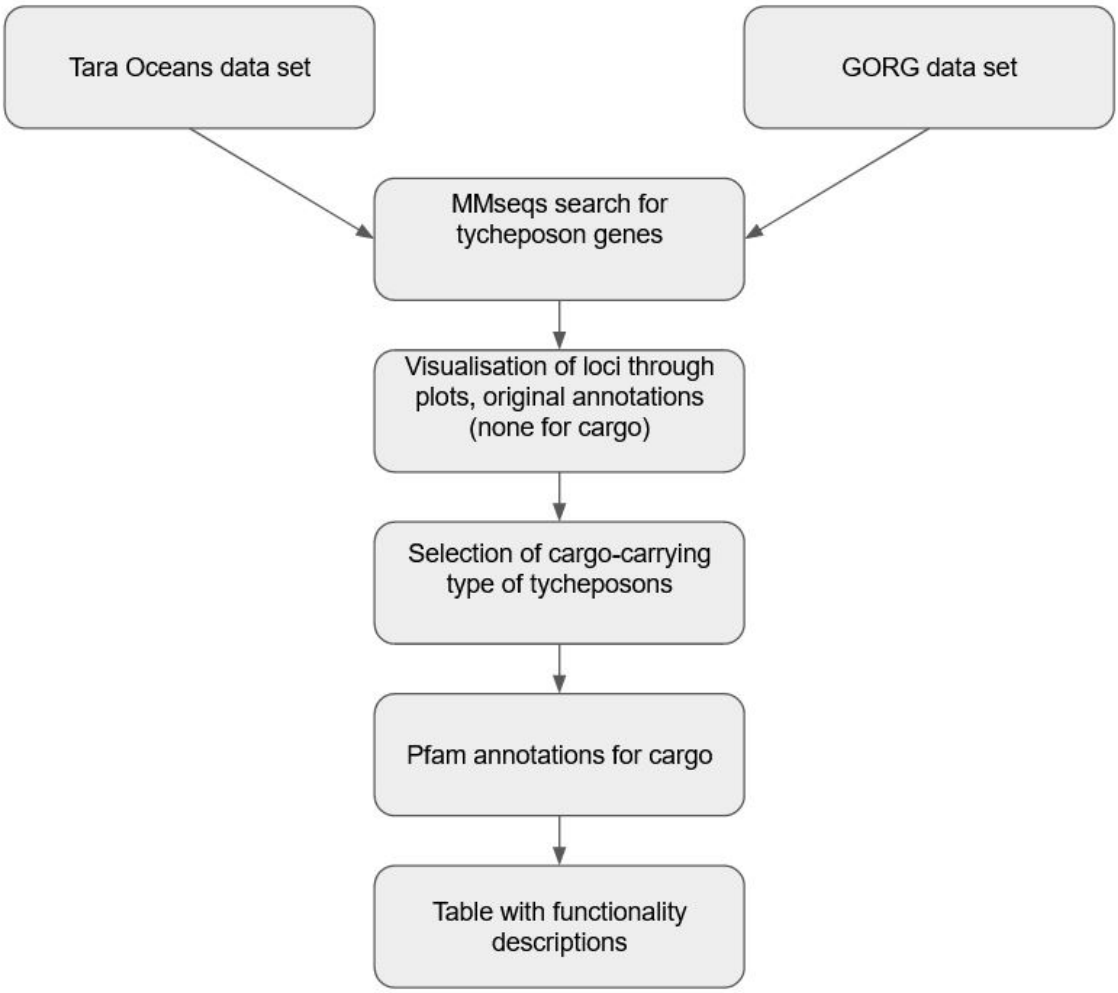


Figure 2: The visualisation and analysis workflow.

4 Results

The result of merging the three tables is a table containing the following main elements: genomic sequence IDs, the IDs of the tycheposon gene hits, the start and the end positions, the modules and the roles, along with several other columns of details used for plotting. This table was taken from the previous project for the Tara Oceans data set, and one was created for the GORG data set as well. In the Tara Oceans results, the preliminary number of genes found was 1537, and upon filtering only 1352 genes remained. In the GORG results, the preliminary number of genes found was 1665639, and upon filtering 1314049 genes remained. The preliminary visualisation of the data was done by computing the loci of interest in a chromosomal context, without further details, to observe tycheposon loci presence in the data set. Then, each locus was plotted in detail, with its gene structure and annotations. Overall, based on the previous research project, the prevalence of tycheposons appears to be quite low on a global oceanic scale than in Cyanobacteria in particular. One observation in the case of the Tara Oceans data set is that some of the loci seem to be composed of multiple tycheposons. One such example can be seen in the first sequence of Appendix 1, which contains several integrases. The Pfam annotations added to the discovered cargo-carrying tycheposons were all taken from the Pfam database [24] and can be found in the supplementary tables [25][26]. The Tara Oceans results contain 1132 annotations, while the GORG results contain 351 annotations.

4.1 Tara Oceans

The plots showing the loci found in the investigated Tara Oceans genomes with their annotated genes can be found in Appendix 2, but an excerpt can be seen in Figure 3. Within the Tara Oceans sequences, 20 loci distributed in 19 genomes with a PICI-like structure and 117 loci distributed in 116 genomes with non-PICI-like structures were discovered. Of these, only 2 genomes had both kinds of tycheposons. This translates to tycheposon presence in 5.12% of the Tara Oceans genomes. Within the discovered loci, 85.4% had a non-PICI-like structure.

Table 1: Cargo of interest - Tara Oceans. Includes the Pfam ID of the genes, their functional annotations, and the number of occurrences in the data set.

Pfam ID	Annotation	Number
PF07862	Nif11 nitrogen-fixing domain	8
PF01032	Fe(3+) dicitrate transport system permease proteins	2
PF01475	Ferric uptake regulator proteins	2
PF02069	Metallothioneins	2
PF04362	Bacterial Fe(2+) trafficking proteins	1
PF04773	FecR, involved in regulation of iron dicitrate transport	1

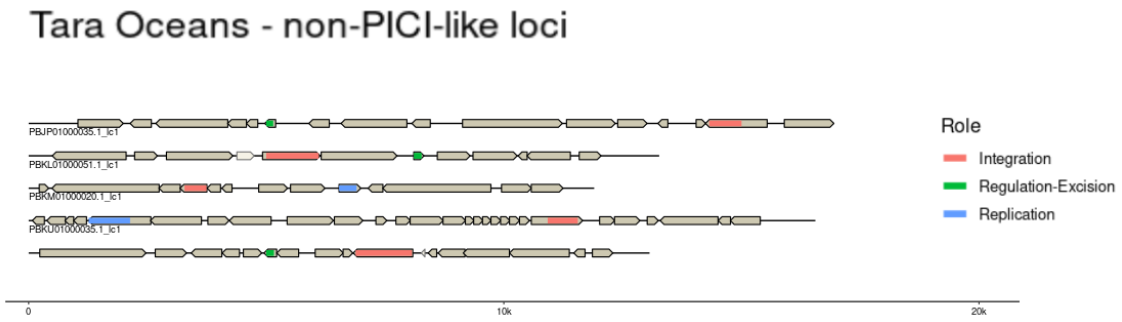


Figure 3: An excerpt of five tycheposon loci from the Tara Oceans plots prior to Pfam annotation. The colors indicate the role of the gene.

Following the Pfam annotation, the most numerous were related to integration, regulation, and replication, characteristic to the tycheposon hallmark genes, as can be seen in Supplementary Table 1. Some virulence-associated factors were discovered as well, for example the virulence-associated protein E, or the inovirus GP2 which plays a crucial role in viral DNA replication. The most abundant cargo seen in the Tara Oceans cargo-carrying tycheposons is a Nif11 domain (Table 1). This is a domain usually found in Cyanobacteria and Proteobacteria, in the Niff11 protein described as nitrogen-fixing in *Azotobacter vinelandii* [27]. Eight instances of this domain have been observed in the Tara Oceans cargo-carrying tycheposons. Other interesting cargoes are the Fe(3+) dicitrate transport system permease proteins FecC and FecD and the FecR protein - involved in regulation of iron dicitrate transport, ferric uptake regulator proteins, metallothioneins, and bacterial Fe(2+) trafficking proteins. Some other notable findings are Beta-lactamases and the Metallo-beta-lactamase superfamily, which are involved in antibiotic resistance, cold-shock proteins, and the SelR domain for coping with oxidative stress.

4.2 GORG

The plots showing the loci found in the investigated GORG genomes with their annotated genes can be found in Appendix 2, but an excerpt can be seen in Figure 4. Within the GORG sequences, only one locus with a PICI-like structure and 40 loci with non-PICI-like structures were discovered, all distributed in different genomes. This translates to tycheposon presence in 0.32% of the GORG genomes. Within the discovered loci, 97.6% had a non-PICI-like structure.

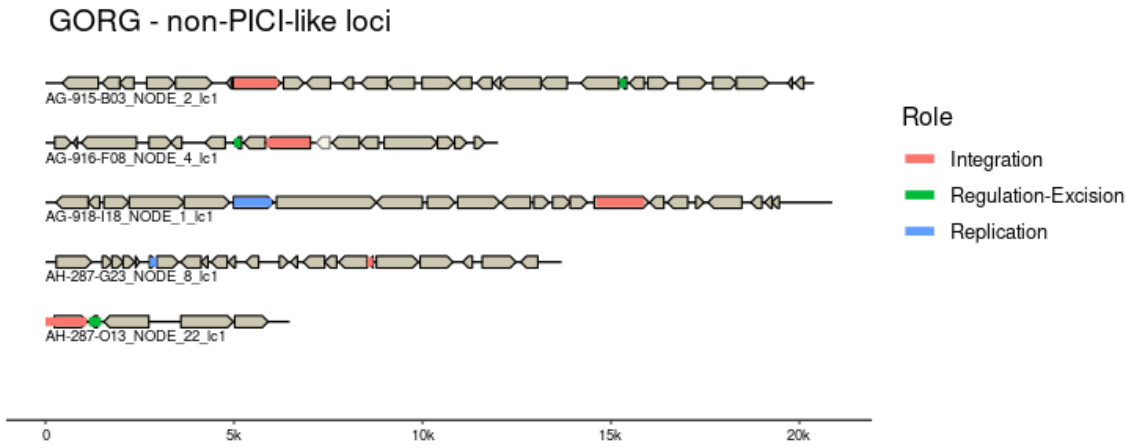


Figure 4: An excerpt of five tycheposon loci from the GORG plots prior to Pfam annotation. The colors indicate the role of the gene.

As in the case of the previously discussed data set, the most numerous annotations in the GORG cargo-carrying tycheposons were related to integration, regulation, and replication, which are characteristic to the tycheposon hallmark genes. The discovered cargoes related to nutrient uptake were the plastocyanin family of copper binding proteins, the FecR protein involved in regulation of iron dicitrate transport, the 4Fe-4S single cluster domain which binds iron-sulfur clusters, and one instance of a Zinc-binding metallo-peptidase family protein. Some interesting cargoes discovered are heat shock proteins, sirtuins which influence processes like aging, transcription, apoptosis, inflammation and stress resistance, energy efficiency and alertness during low-calorie situations, flavoproteins and the SelR domain involved, among others, in resisting oxidative stress, a PhoH-like protein involved in phosphate starvation response, Poly(hydroxyalcanoate) granule associated protein (phasin) functioning as an intracellular carbon and energy reserve material.

Table 2: Cargo of interest - GORG. Includes the Pfam ID of the genes, their functional annotations, and the number of occurrences in the data set.

Pfam ID	**Annotation**	**Number**
PF04773	FecR protein - involved in regulation of iron dicitrate transport	1
PF00127	Plastocyanin family of copper binding proteins	1
PF13353	4Fe-4S single cluster domain - bind iron-sulfur clusters	1
PF13582	Zinc-binding metallo-peptidase family	1
PF00118	Heat shock proteins	2
PF02146	Sirtuins - stress resistance, management of low-calorie situations, etc.	2
PF02441	Flavoproteins - management of oxidative stress, etc.	2
PF01641	SelR domain - management of oxidative stress	1
PF05597	Poly(hydroxyalcanoate) granule associated protein - energy reserve material	1

5 Discussion and Conclusions

This study investigated the ecological relevance of the functional cargo of tycheposons. The number of tycheposons discovered in these two data sets was quite low, which could be explained in several ways. Firstly, the sequencing methods used for both data sets are inherently prone to exclude small mobile genetic elements. Furthermore, the GORG data set contains incomplete sequences to begin with, which could be another limitation when searching for the small modules which the tycheposon hallmark genes consist of. A further explanation could be the search method in itself. Due to time limitations, the search was based only on the hallmark genes discovered in the original study, which were initially found in Prochlorococcus. However, in other bacteria tycheposons could look quite different from what they have been determined to be so far. A more useful search would be a recursive one where the newly discovered tycheposons, especially those found outside of the Cyanobacteria phylum, are added to the reference hallmark genes set to increase the sensitivity of the search.

In the Tara Oceans data set, one important analysis pitfall which was also observed in the Cyanorak set in the previous project was that, in some cases, it appears that several neighbouring tycheposons were clustered together into one locus, so they were counted as only one element. This is due to the method through which a locus was defined, with no upper limit to the presence of one type of gene. While this is something that could be rectified in future research, during this study the time limitations did not allow for it. This occurrence is still relevant to analyze: the fact that two or more structurally intact tycheposons are located one next to the other within a genome could suggest that the integrases in their composition target the same tRNA. An aspect that could be investigated is whether these neighbouring tycheposons have similar or different functions. This could not done in this study because of the large number of annotations, but it could be a topic of interest in future research. Furthermore, the presence of more tycheposons in neighbouring positions could also mean that they are part of a genomic island with a relatively sustained influx of these elements, which might be the case in an environmental patch with a more dense bacterial population, where HGT is a frequent occurrence. Although this hypothesis could not be tested during this study due to time limitations, it is still a relevant aspect to investigate in future research by conducting a tycheposon biogeography study.

Notably, in the cargo-carrying, non-PICI-like tycheposons analyzed, plenty of virulence- and phage-related factors were discovered, for example the prophage CP4-57 regulatory protein, of which many instances were found in both data sets [25][26]. Since these loci lack the packaging-interference module and were thus not included in the PICI-like category during filtering, they are likely relics of phage parasite tycheposons which were incompletely integrated into the host genomes, or transferred so long ago that some components were lost. The loci containing this type of cargo but no typical packaging-interference module are not of the cargo-carrying category

and should not be included in it, but for the purpose of this study it was unnecessary to filter them out of the results, because it would not have contributed to answering the research question.

As expected, some limiting nutrient uptake cargoes have been found in both of the data sets. Nitrogen uptake cargo has only been discovered in the Tara Oceans data set, in which eight instances of the Nif11 nitrogen-fixing domain were observed. With nitrogen being the fourth most abundant element in organic matter, while also among the most limiting nutrients in the ocean [28], it is clear that this type of cargo confers a tremendous advantage and growth opportunity to the host marine bacteria. The oceans are a key component of the biosphere, hosting a variety of biogeochemical processes influencing both the atmosphere and the land [28]. The nitrogen cycle is one of these important processes, in which nitrogen in different chemical forms is exchanged between the atmosphere and the terrestrial and marine ecosystems [29]. In this cycle, the ocean is involved in both nitrogen fixation - through microorganisms performing this function - and denitrification into the atmosphere. With a tycheposon with a nitrogen fixation ecological function, the host bacteria can partake into the nitrogen cycle and contribute to the proper functioning of the ecosystem, while also benefiting from the ability to take up one of the most limiting nutrients.

Another type of nutrient uptake cargo that was found in both sets is cargo related to iron uptake. Iron is a micronutrient known to be crucial in many marine biogeochemical cycles and in the maintenance of pelagic ecosystems [30]. Iron is one of the most important resources which influence the extent and patterns of primary productivity in the ocean, being a co-factor in enzymes involved in basic life-sustaining processes such as respiration, photosynthesis and nitrogen fixation [31]. Through its role in these processes, iron is also associated with the marine part of the carbon cycle, in which the marine biological pump absorbs carbon from the atmosphere and land runoff and acts as a carbon sink [32]. Due to the low solubility of iron which limits the attainability of this element to marine microorganisms [17], iron uptake is a notably useful function for tycheposon cargo, which is likely why this type of cargo genes are so prevalent in the two explored data sets.

Metallothioneins were also found in the Tara Oceans data set. They are small proteins which bind heavy metals such as copper, zinc and nickel [33]. In the GORG data set one instance of zinc-binding metallo-peptidase family protein was found. While zinc is not one of the major limiting nutrients in the ocean, it is still a necessary and quite scarce trace-nutrient outside of high-latitude regions [34]. It has been found that zinc might play a role in the composition and productivity of phytoplankton communities [34]. Another cargo related to essential heavy metal uptake was also found in the latter data set - Plastocyanin copper binding proteins. This type of cargo seems to be involved in the uptake of essential heavy metals - rather than non-essential and toxic ones such as mercury or cadmium. The slight differences between the sets could arise from the difference in the sampling areas which the bacteria were taken from. While the Tara expedi-

tion covered all types of waters, from tropical areas to cold zones near the poles, the GORG data comes only from the tropical and sub-tropical regions, so they cover fewer possible environmental conditions.

An interesting finding in both of the data sets is the SelR domain, which is responsible, among other functions, for oxidative stress resistance. Aside from this domain, several others were found with the same function (i.e. flavoproteins), suggesting that this type of functional cargo might be a rather wide-spread occurrence. Oxidative stress is a phenomenon occurring in the case of an imbalance between the production and accumulation of oxygen reactive species in cells, and the capacity of a biological system to detoxify these reactive products [35]. Some environmental stressors which could cause this type of phenomenon are pollutants, UV radiation, ionizing radiation, and heavy metals. Oxidative stress resistance is not an uncommon function for mobile genetic elements, as it can also be seen in the integrative conjugative element ICE-βox, which was proven in the lab to confer oxidative stress resistance to *Legionella pneumophila* against antibiotics such as penicillin and oxacillin, but also against bleach [2]. This cargo is likely found in Cyanobacterial phyla, since it is known that, due to their role as oxygenic phototrophs, Prochlorococcus and Synechococcus often encounter elevated levels of oxidative stress from UV exposure [36], and would thus benefit greatly from a defense mechanism against this nocive factor. Still, this kind of cargo, along with the cargo involved in heavy metal binding, could make tycheposons a great adaptation tool for other hosts as well, especially the current situation where anthropogenic ocean pollution is becoming an increasingly more significant problem, which could promote their spread even outside Cyanobacteria.

Similarly, another type of cargo found in both data sets is related to temperature shock, more precisely cold-shock proteins in Tara Oceans and heat-shock proteins in GORG, which are responsible for maintaining cellular proteins integrity during environmental changes [37]. These functions, too, confer better adaptability to the host, since there are plenty of areas in the ocean where temperatures vary considerably on a regular basis. This adaptation could also be especially relevant in the context of the current climate change situation, which causes more extreme temperature changes in the oceanic waters as well as in the atmosphere, especially at the surface, in the euphotic zone where most microorganisms reside. These changes lead to alterations of the environment to which bacteria need to adapt, so MGEs with temperature shock management functions would confer a great advantage.

The discovery of these types of cargoes suggests that cargo-carrying tycheposons may have a more diverse array of functions than previously known, possibly providing adaptation to anthropogenic and climate-change-related issues as well, aside from their previously known functions in limiting nutrient uptake. This could be an evolutionary adaptation in response to the changing environmental conditions in the recent decades, which continue to affect the ocean greatly, both

in terms of temperature variation and of pollution from diverse human activities. This type of cargo might become more and more relevant as these changes advance, so it would be expected for tycheposons carrying it to become a more common coping mechanism in the affected marine bacterial populations. Furthermore, in future research, it would be relevant to explore tycheposon prevalence in marine areas already known to have a problem with pollution, for example the heavily polluted Mediterranean Sea [38] or the Gulf of Mexico Dead Zone [39]. In this kind of locations, marine bacteria need to adapt to the harsh conditions, so tycheposons with environmental stress relief functions could be more useful, and thus also more prevalent. In such regions, it would be expectable to find more tycheposons of this type, which could bring to light the range of functions they can perform in adaptation to detrimental anthropogenic effects on the environment.

Tycheposons with limiting nutrient uptake cargo could also be found in these areas, since they are usually quite rich in nutrients such as nitrogen and phosphorous coming from the shores, while also limited in oxygen. It would also be relevant to search for tycheposons and explore their functions in bacteria from regions of the oceans that are notoriously nutrient-scarce, such as oxygen minimum zones. In these regions adaptive microbial metabolism is extremely important [40], and thus the horizontally transferred material in the genomic islands of the inhabitants likely has functions in limiting nutrient uptake. These areas are relatively unexplored, but taking into consideration the utility that tycheposons may have for organisms living in such environments, it is likely that they are present there. Researching their presence and function in these environments may contribute to our understanding of how they have shaped microbial adaptability to harsh living conditions. Furthermore, considering that previous research has found that, besides adaptation advantages, horizontally transferred material can also increase the efficiency of resource usage for bacterial reproduction [41], it would also be relevant to explore whether some tycheposons, likely outside of Prochlorococcus, also have such an effect on their host.

Overall, it seems that cargo-carrying tycheposon genes may have quite a wide variety of functions, all of which facilitate adaptation to the environment, confer resistance to external stressors such as harsh environmental conditions, pollution, and climate change effects, and even introduce their hosts in some of the main nutrient cycles taking place in the ocean. These advantageous effects they have on the bacteria which carry them promotes their dispersal through populations on a large scale. Still, as they are relatively newly discovered MGEs, there is a lot of research to be done before we can untangle all aspects of tycheposon ecology and fully understand their implications in the marine bacterial world.

References

[1] Burmeister, A. R. (2015). Horizontal gene transfer. Evolution, Medicine, and Public Health, 2015(1), 193–194. https://doi.org/10.1093/emph/eov018

[2] Flynn, K. J., & Swanson, M. S. (2014). Integrative conjugative element ice-βox confers oxidative stress resistance to legionella pneumophila in vitro and in macrophages. Mbio, 5(3), 01091–14. https://doi.org/10.1128/mBio.01091-14

[3] Krishnapillai, V. (1996). Horizontal gene transfer. Journal of Genetics, 75(2), 219–232. https://doi.org/10.1007/BF02931763.

[4] Garcia-Aljaro, C., Balleste, E., Muniesa, M. (2017). Beyond the canonical strategies of horizontal gene transfer in prokaryotes, Current Opinion in Microbiology, 38, 95-105, ISSN 1369-5274, doi: https://doi.org/10.1016/j.mib.2017.04.011.

[5] Levin, H., Moran, J., Malik, H., Craig, N., Bourque, G., Dubnau, J., Slotkin, R., Flasch, D., Gunderson, K., Feschotte Cedric, Peters, J., & Singh, P. (2014). Mobile genetic elements and genome evolution 2014. Mobile Dna, 5(1), 1–10. https://doi.org/10.1186/1759-8753-5-26

[6] Kosecka-Strojek, M., Buda, B., Miedzobrodzki, J. (2018). Pet-To-Man Travelling Staphylococci: a World in Progress. Academic Press. ch. 2, 11-24, ISBN 9780128135471, https://doi.org/10.1016/B978-0-12-813547-1.00002-9.

[7] Summers, D. K. (1996). "Chapter 1 – The Function and Organization of Plasmids". The biology of plasmids. Blackwell Science. https://doi.org/10.1002/9781444313741

[8] Alfred, F.-S., Roser Martinez-Rubio, Rezheen, F. A., John, C., Robert, D., & Jose R. Penades. (2018). Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. The Isme Journal, 2114-2128, 2114–2128. https://doi.org/10.1038/s41396-018-0156-3

[9] Hackl, T. et al. (2020). Novel integrative elements and genomic plasticity in ocean ecosystems. BioRxiv, doi: https://doi.org/10.1101/2020.12.28.424599

[10] Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., & Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of prochlorococcus. Science, 311(5768), 1768–1770.

[11] Penades, J. R., & Christie, G. E. (2015). The phage-inducible chromosomal islands: a family of highly evolved molecular parasites. Annual Review of Virology, 2(1), 181–201. https://doi.org/10.1146/annurev-virology-031413-085446.

[12] Chen, J., Carpena, N., Quiles-Puchalt, N., Ram, G., Novick, R. P., & Penades J. R. (2015). Intra- and inter-generic transfer of pathogenicity island-encoded virulence genes by cos phages. The Isme Journal, 9(5), 1260–3. https://doi.org/10.1038/ismej.2014.187.

[13] Hackl, T., Duponchel, S., Barenhoff, K., Weinmann, A. & Fischer, M. G. (2020). Endogenous virophages populate the genomes of a marine heterotrophic flagellate. BioRxiv, doi:10.1101/2020.11.30.404863.

[14] Coale, K. H. (1991). Effects of iron, manganese, copper, and zinc enrichments on productivity and biomass in the subarctic pacific. Limnology and Oceanography, 36(8), 1851–1864. https://doi.org/10.4319/lo.1991.36.8.1851

[15] Phosphorus. [Online]. Available: https://ods.od.nih.gov/factsheets/Phosphorus-HealthProfessional/ (Accessed on: 24-06-2021).

[16] Aczel, M. R. (2019). What is the nitrogen cycle and why is it key to life? Frontiers for Young Minds, 7. https://doi.org/10.3389/frym.2019.00041

[17] Ferreira, F., & Straus, N. A. (1994). Iron deprivation in cyanobacteria. Journal of Applied Phycology, 6(2), 199–210. https://doi.org/10.1007/BF02186073

[18] European Nucleotide Archive. [Online]. Available: https://www.ebi.ac.uk/ena/browser/text-search?query=tara (Accessed on: 26-04-2021).

[19] Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., & Stepanauskas, R. (2019). Charting the complexity of the marine microbiome through single-cell genomics. Cell, 179(7), 1623–1635. https://doi.org/10.1016/j.cell.2019.11.017

[20] Nawy, T. (2014). Single-cell sequencing. Nature Methods, 11(1), 18–18.

[21] Supplementary code and data for mobile genetic elements in Prochlorococcus. [Online]. Available: https://github.com/thackl/pro-tycheposons (Accessed on: 20-05-2021).

[22] MMseqs2 User Guide. [Online]. Available: https://mmseqs.com/latest/userguide.pdf (Accessed on: 26-04-2021).

[23] SeqKit - a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. [Online]. Available: https://bioinf.shenwei.me/seqkit/ (Accessed on: 28-04-2021).

[24] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., & Bateman, A. (2010). The pfam protein families database

[25] Jecu, Diana (2021): Supplementary Table 1 - Tara Oceans.xlsx. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14852154.v1

[26] Jecu, Diana (2021): Supplementary Table 2 - GORG.xlsx. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14852151.v1

[27] Jacobson, M. R., Brigle, K. E., Bennett, L. T., Setterquist, R. A., Wilson, M. S., Cash, V. L., Beynon, J., Newton, W. E., & Dean, D. R. (1989). Physical and genetic map of the major nif gene cluster from azotobacter vinelandii. Journal of Bacteriology, 171(2), 1017–27.

[28] Zehr, J. P., & Kudela, R. M. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. Annual Review of Marine Science, 3, 197–225. https://doi.org/10.1146/annurev-marine-120709-142819

[29] Fowler, D., Coyle, M., Skiba, U., Sutton, M. A., Cape, J. N., Reis, S., Sheppard, L. J., Jenkins, A., Grizzetti, B., Galloway, J. N., Vitousek, P., Leach, A., Bouwman, A. F., Butterbach-Bahl, K., Dentener, F., Stevenson, D., Amann, M., & Voss, M. (2013). The global nitrogen cycle in the twenty-first century. Philosophical Transactions of the Royal Society of London. Series B, 368(1621).

[30] Boyd, P. W., Strzepek, R. F., Ellwood, M. J., Hutchins, D. A., Nodder, S. D., Twining, B. S., & Wilhelm, S. W. (2015). Why are biotic iron pools uniform across high- and low-iron pelagic ecosystems? Global Biogeochemical Cycles, 29(7), 1028–1043. https://doi.org/10.1002/2014GB005014

[31] Tagliabue, A., Bowie, A. R., Boyd, P. W., Buck, K. N., Johnson, K. S., & Saito, M. A. (2017). The integral role of iron in ocean biogeochemistry. Nature, 543(7643), 51–59. https://doi.org/10.1038/nature21058

[32] Sigman, D.M. & Haug, G.H. (2006). The biological pump in the past. In: Treatise on Geochemistry, 6, (ed.). Pergamon Press, 491-528.

[33] Stillman, M. J. (1995). Metallothioneins. Coordination Chemistry Reviews, 144(Com), 461–511.

[34] Middag, R., Baar, H. J. W., & Bruland, K. W. (2019). The relationships between dissolved zinc and major nutrients phosphate and silicate along the geotraces ga02 transect in the west atlantic ocean. Global Biogeochemical Cycles, 33(1), 63–84. https://doi.org/10.1029/2018GB006034

[35] Gabriele, P., Natasha, I., Mariapaola, C., Giovanni, P., Federica, M., Vincenzo, A., Francesco, S., Domenica, A., & Alessandra, B. (2017). Oxidative stress: harms and benefits for human health. Oxidative Medicine and Cellular Longevity, 2017. https://doi.org/10.1155/2017/8416763

[36] Mella-Flores, D., Six, C., Ratin, M., Partensky, F., Boutte, C., Le Corguille G, Marie, D., Blot, N., Gourvil, P., Kolowrat, C., & Garczarek, L. (2012). Prochlorococcus and synechococcus have evolved different adaptive mechanisms to cope with light and uv stress. Frontiers in Microbiology, 3, 285–285.

[37] Burdon, R. H. (1988). The heat shock proteins. Endeavour, 12(3), 133–138. https://doi.org/10.1016/0160-9327(88)90134-2

[38] Cheryl, H. (2017). Gulf of mexico dead zone is largest ever. C&En Global Enterprise, 95(32), 15–15. https://doi.org/10.1021/cen-09532-govcon1

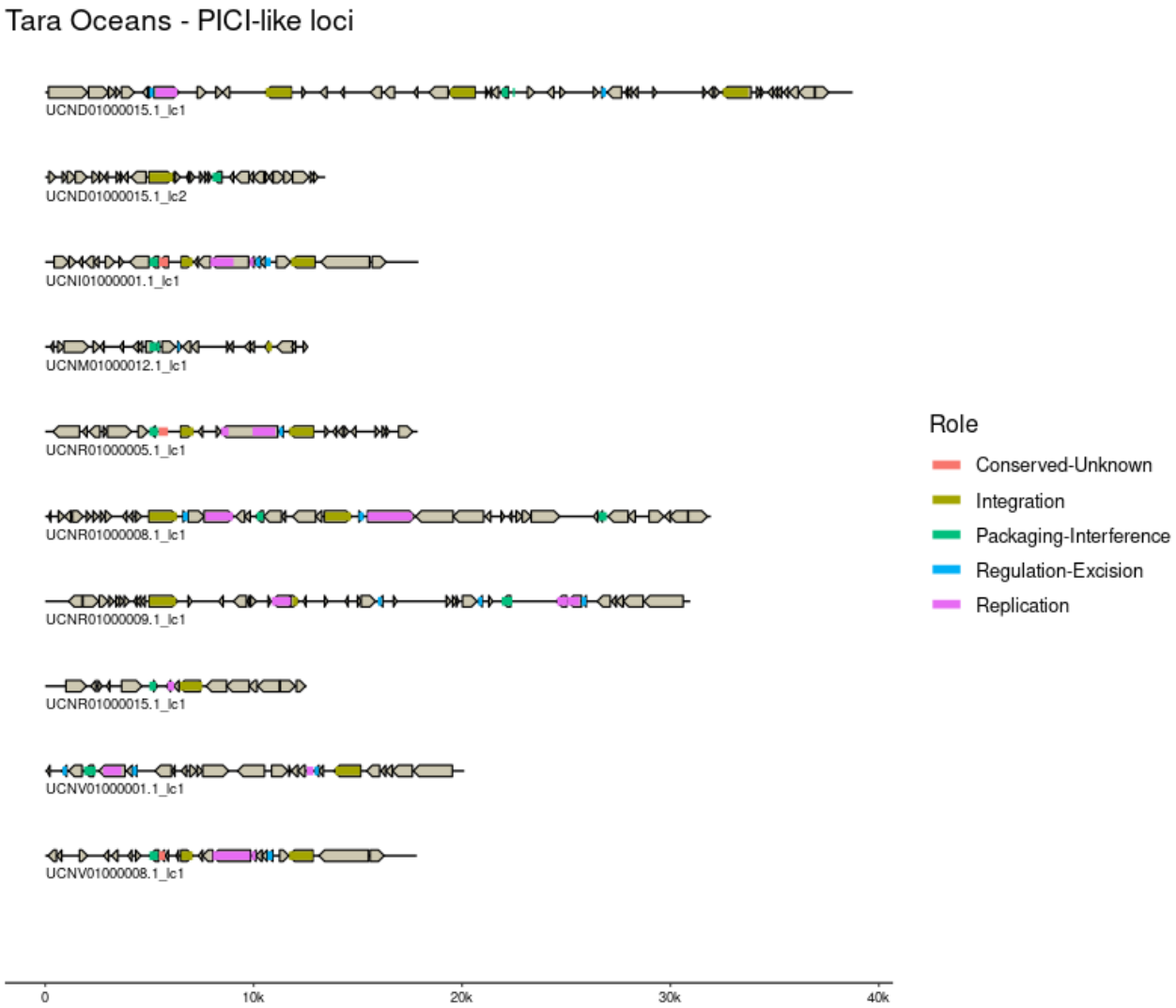
[39] Osterberg, C., & Keckes, S. (1977). The state of pollution of the mediterranean sea. Ambio, 6(6), 321–326.

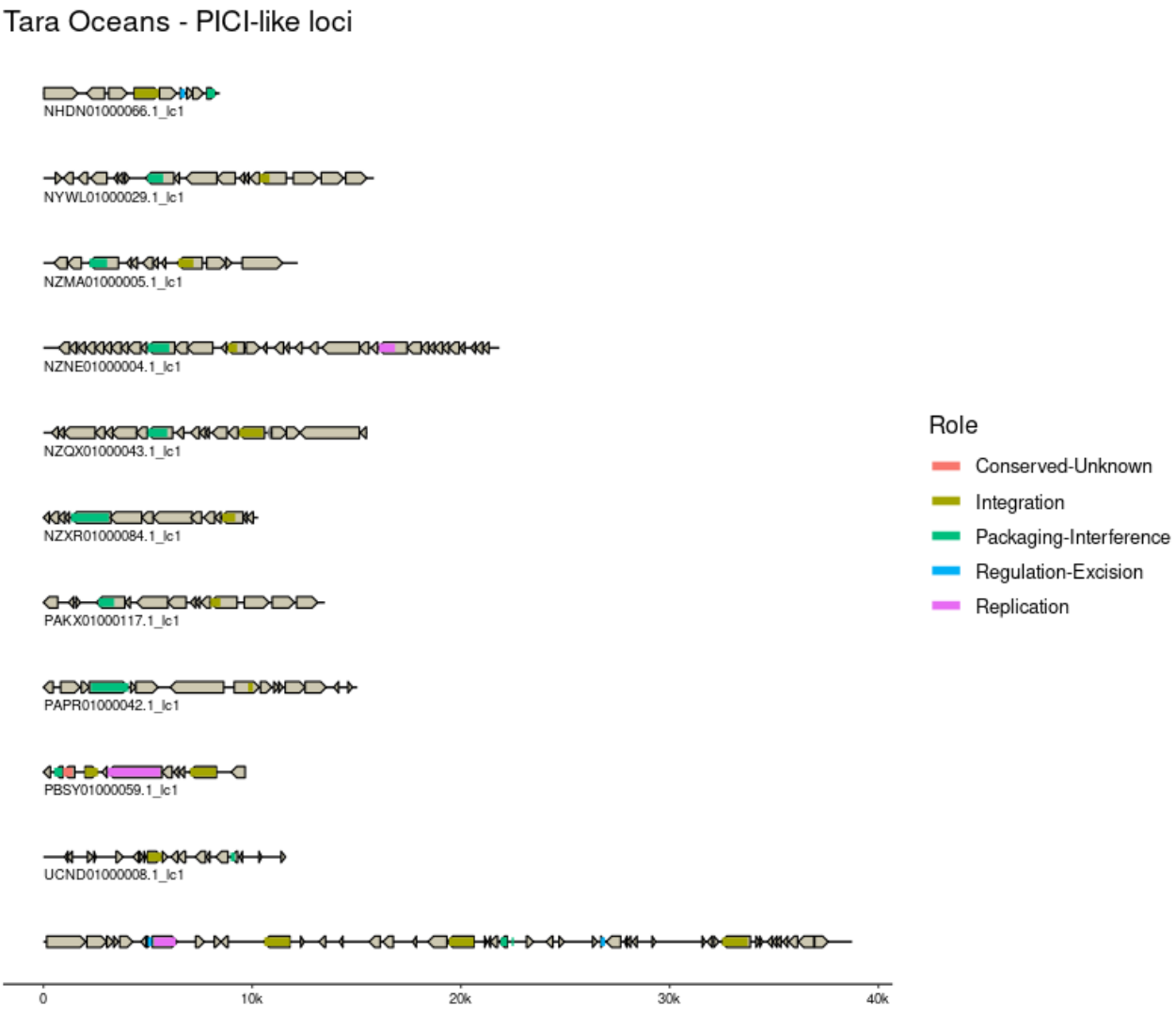
[40] Cheryl-Emiliane, T. C., Danielle, M. W., Richard, A. W. I. I. I., Steven, J. H., & Curtis, A. S. (2015). Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. Frontiers in Microbiology, (2015). https://doi.org/10.3389/fmicb.2015.00265

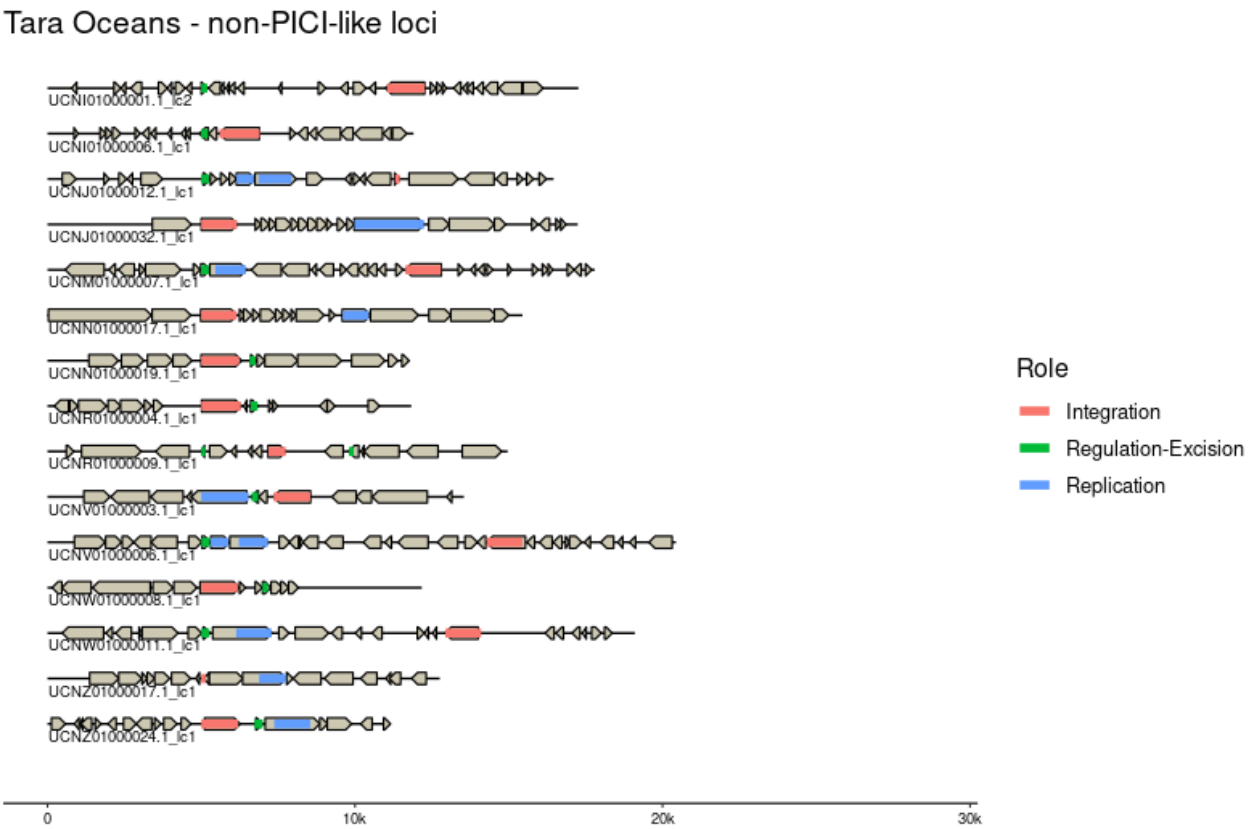
[41] Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovacs, K., Mehi, O., Baliko, G., Szappanos, B., Gyorfy, Z., Feher, T., Bogos, B., Blattner, F. R., Paal, C., Posfai, G., & Papp, B. (2016). Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. Molecular Biology and Evolution, 33(5), 1257–69. https://doi.org/10.1093/molbev/msw009

6 Appendices

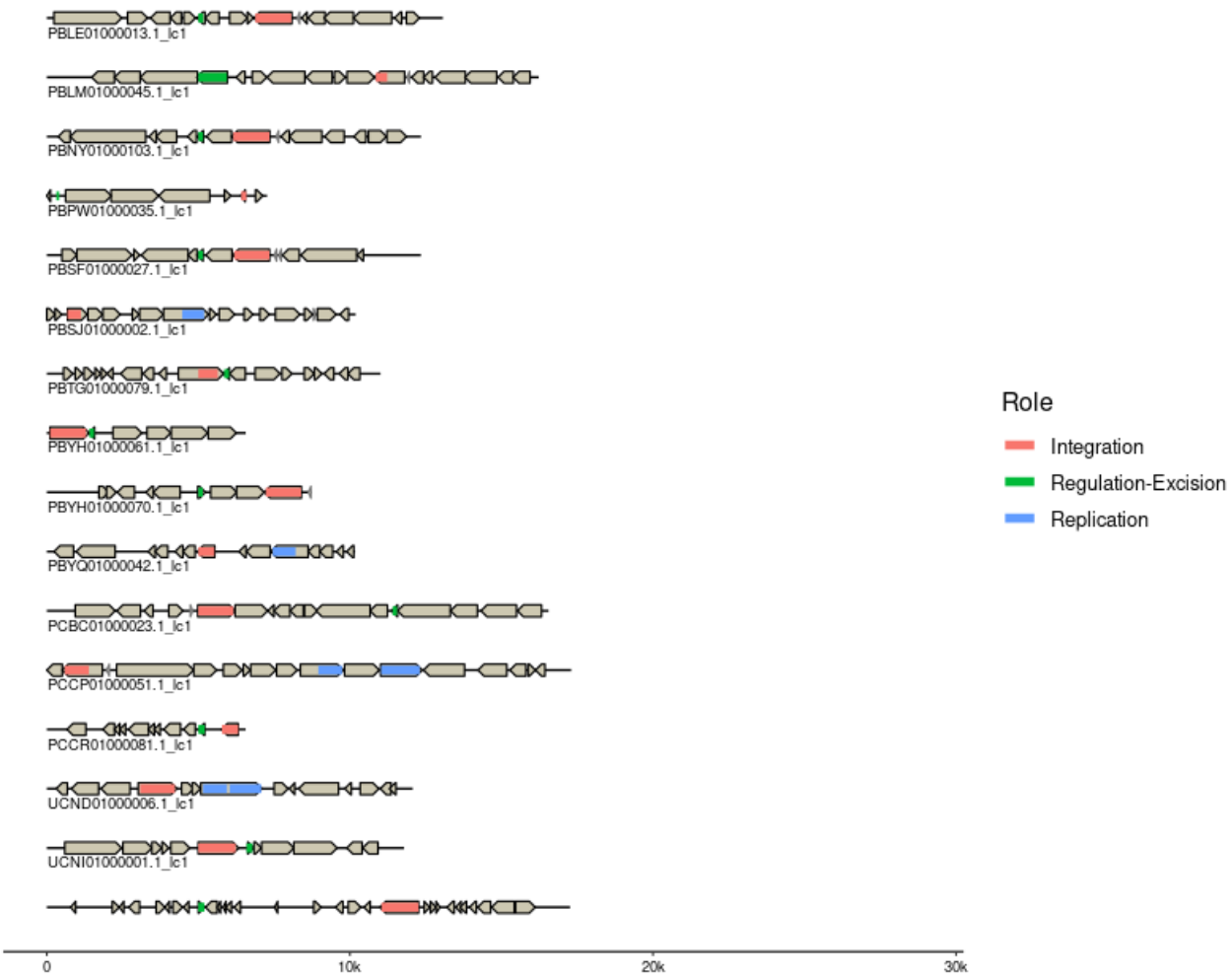
6.1 Appendix 1 - Tara Oceans

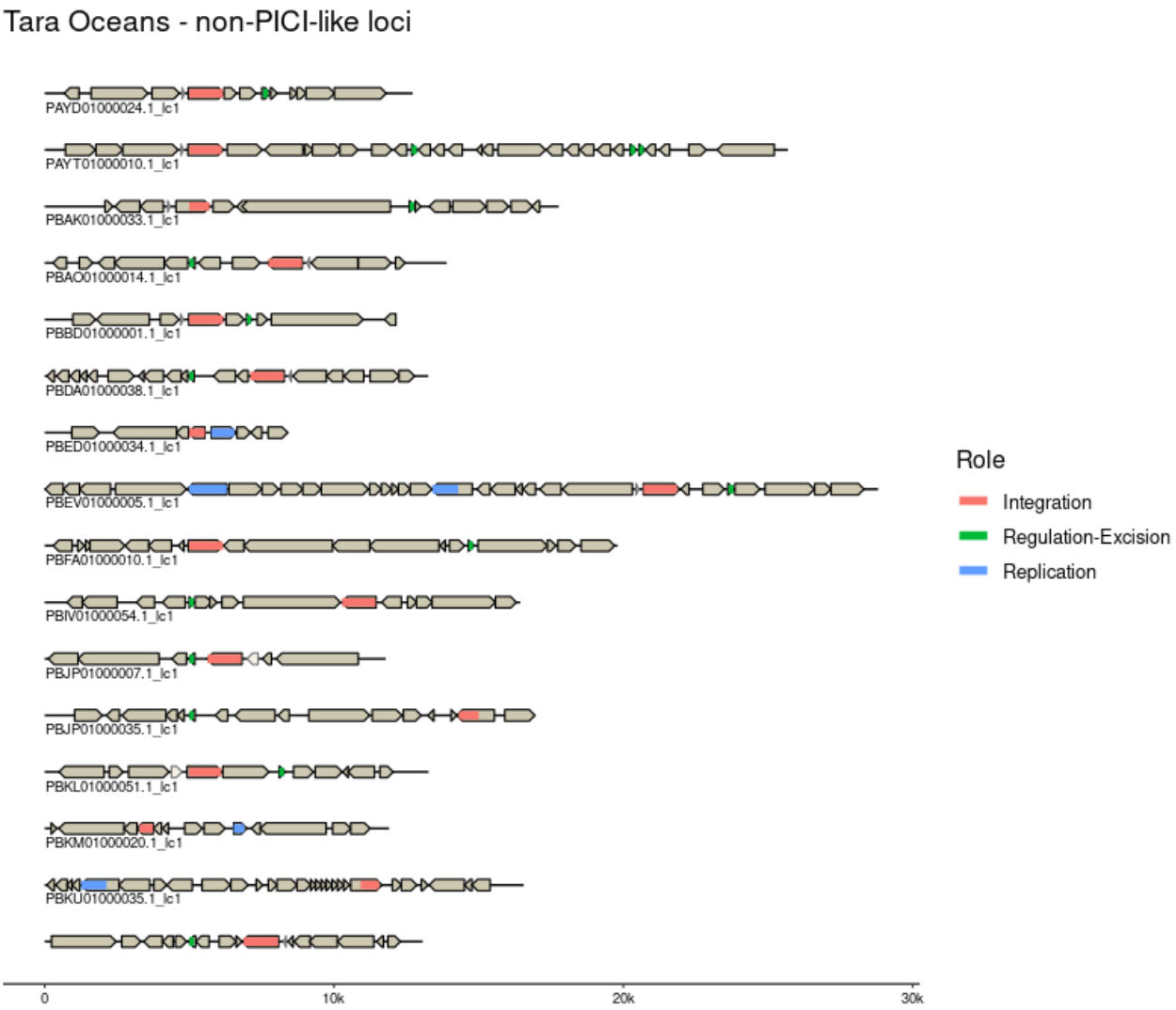


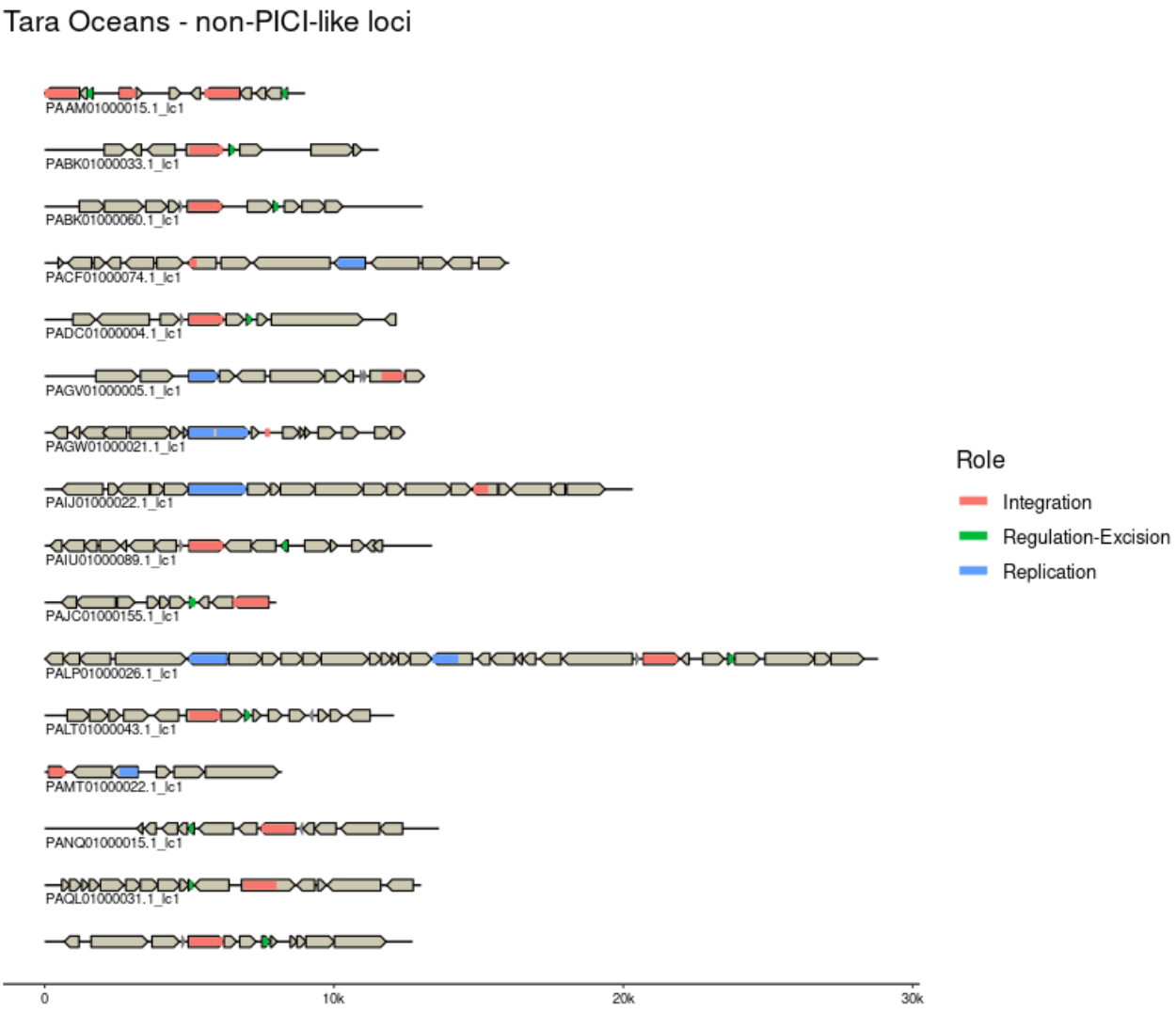

Tara Oceans - PICI-like loci
NHDN01000066.1_lc1
NYWL01000029.1_lc1
NZMA01000005.1_lc1
NZNE01000004.1_lc1
NZQX01000043.1_lc1
NZXR01000084.1_lc1
PAKX01000117.1_lc1
PAPR01000042.1_lc1
PBSY01000059.1_lc1
UCND01000008.1_lc1
0
10k
20k
30k
40k
Role
Conserved-Unknown
Integration
Packaging-Interference
Regulation-Excision
Replication


Tara Oceans - non-PICI-like loci
UCNI01000001.1_lc2
UCNI01000006.1_lc1
UCNJ01000012.1_lc1
UCNJ01000032.1_lc1
UCNM01000007.1_lc1
UCNN01000017.1_lc1
UCNN01000019.1_lc1
UCNR01000004.1_lc1
UCNR01000009.1_lc1
UCNV01000003.1_lc1
UCNV01000006.1_lc1
UCNW01000008.1_lc1
UCNW01000011.1_lc1
UCNZ01000017.1_lc1
UCNZ01000024.1_lc1
0
10k
20k
30k
Role
Integration
Regulation-Excision
Replication

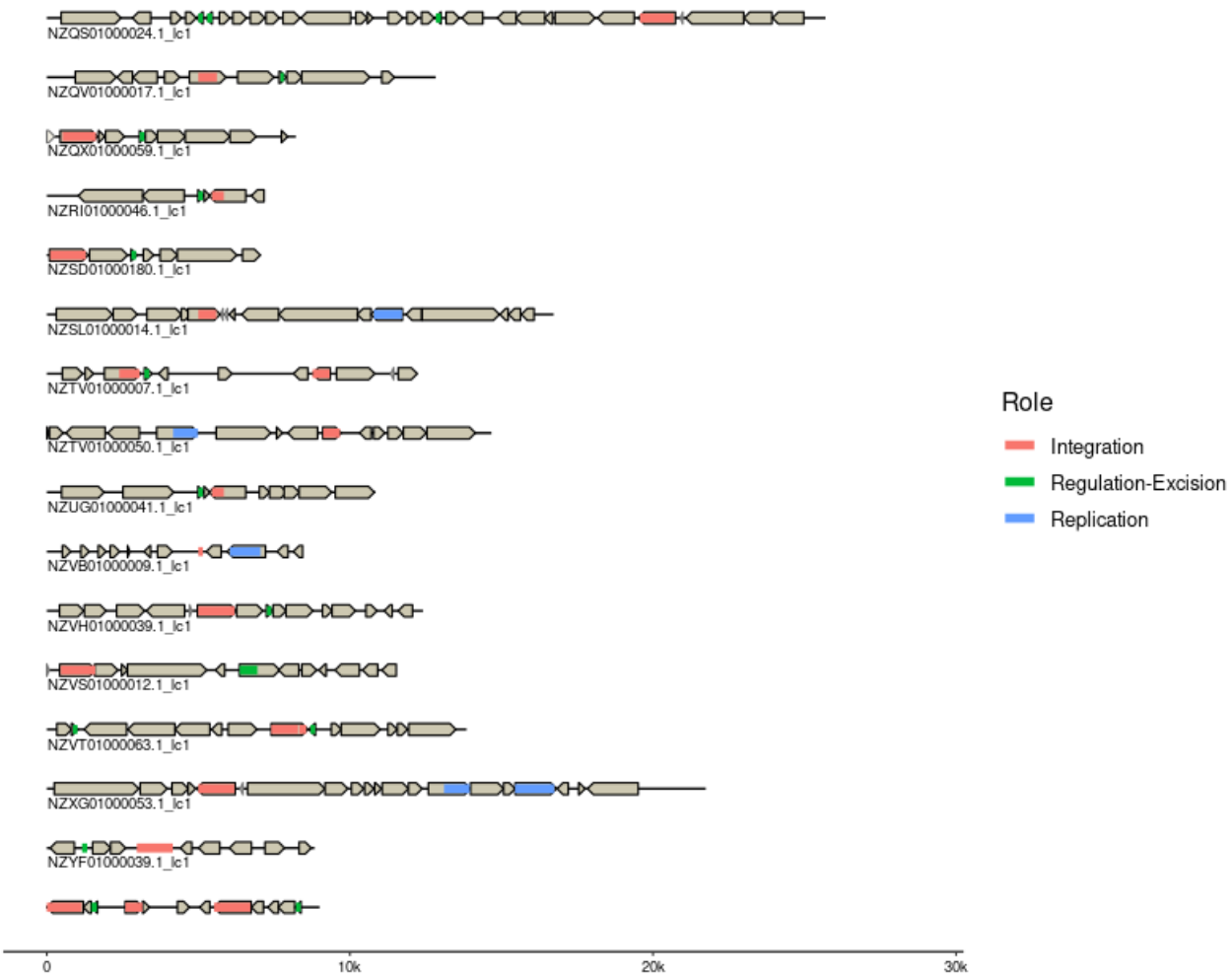
Tara Oceans - non-PICI-like loci



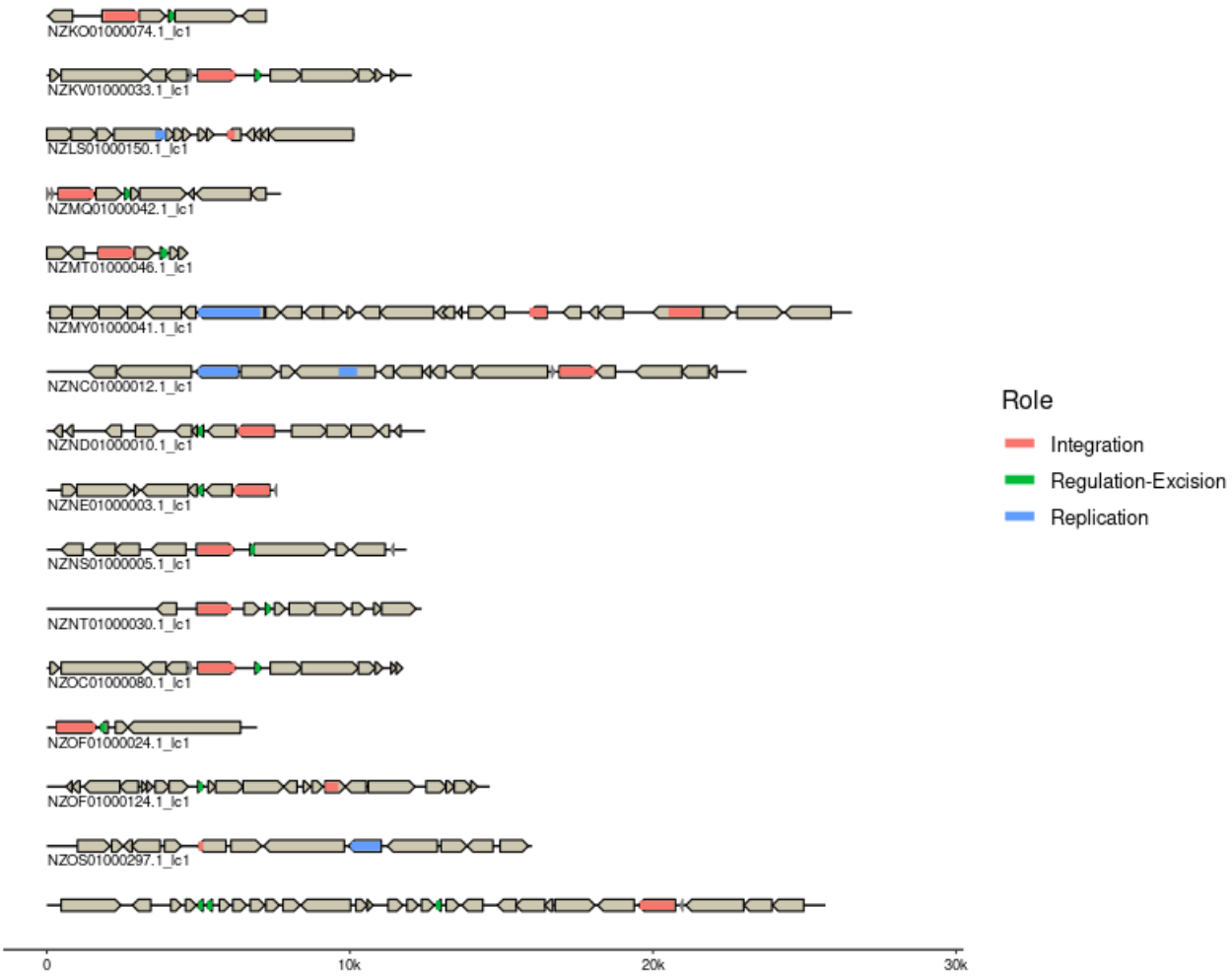

Tara Oceans - non-PICI-like loci
PAYD01000024.1_lc1
PAYT01000010.1_lc1
PBAK01000033.1_lc1
PBAO01000014.1_lc1
PBBD01000001.1_lc1
PBDA01000038.1_lc1
PBED01000034.1_lc1
PBEV01000005.1_lc1
PBFA01000010.1_lc1
PBIV01000054.1_lc1
PBJP01000007.1_lc1
PBJP01000035.1_lc1
PBKL01000051.1_lc1
PBKM01000020.1_lc1
PBKU01000035.1_lc1
0
10k
20k
30k
Role
Integration
Regulation-Excision
Replication

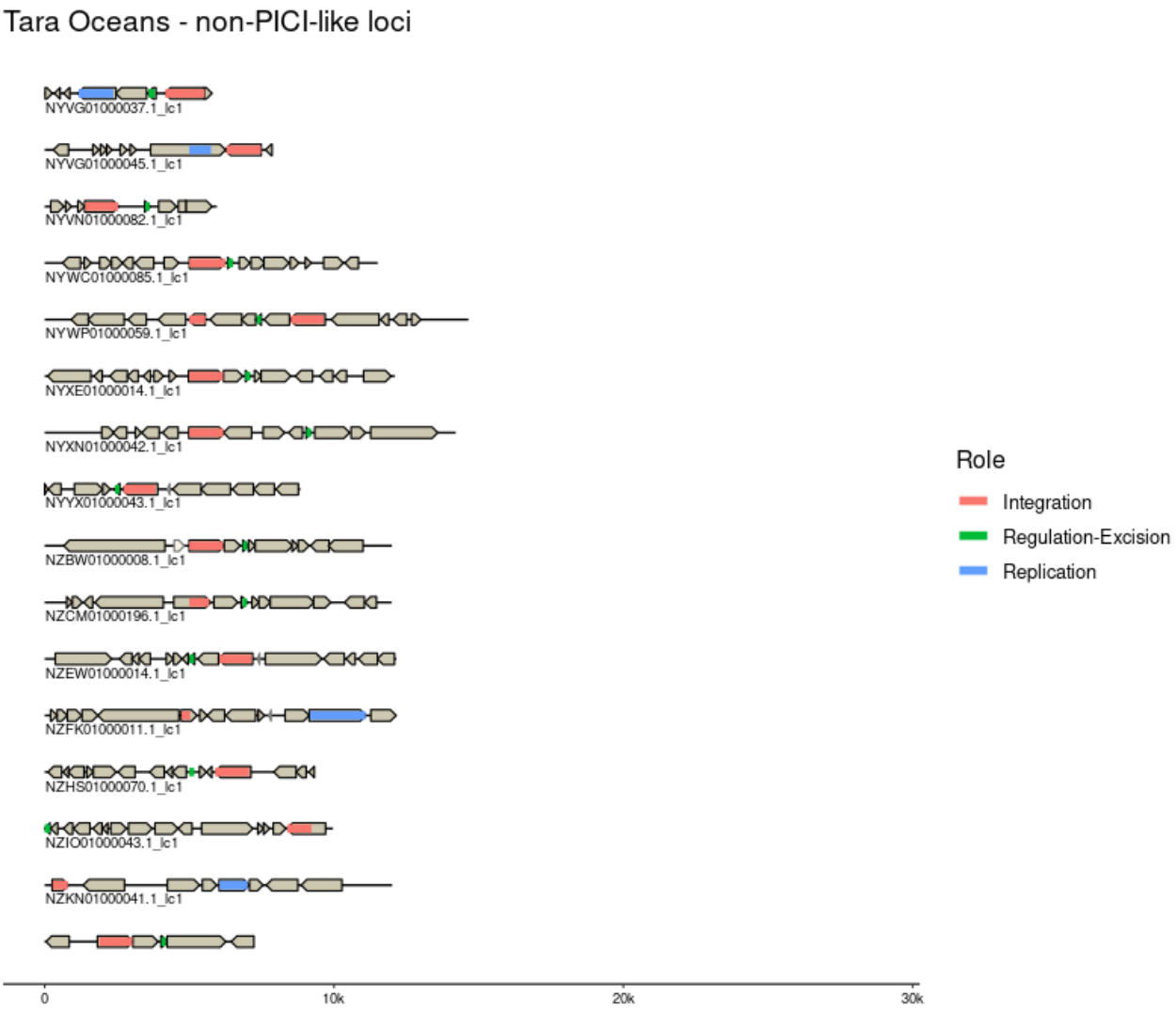

Tara Oceans - non-PICI-like loci
PAAM01000015.1_lc1
PABK01000033.1_lc1
PABK01000060.1_lc1
PACF01000074.1_lc1
PADC01000004.1_lc1
PAGV01000005.1_lc1
PAGW01000021.1_lc1
PAIJ01000022.1_lc1
PAIU01000089.1_lc1
PAJC01000155.1_lc1
PALP01000026.1_lc1
PALT01000043.1_lc1
PAMT01000022.1_lc1
PANQ01000015.1_lc1
PAQL01000031.1_lc1
0
10k
20k
30k
Role
Integration
Regulation-Excision
Replication

Tara Oceans - non-PICI-like loci

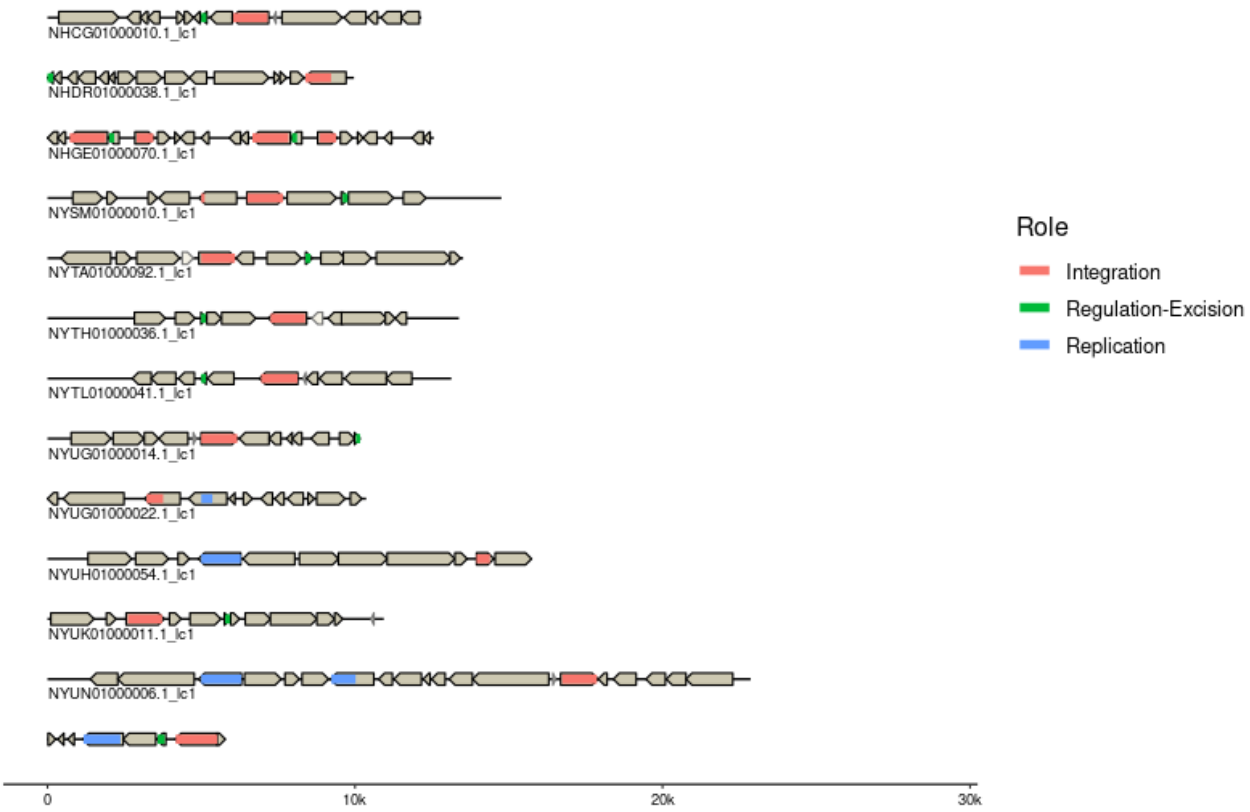


Tara Oceans - non-PICI-like loci

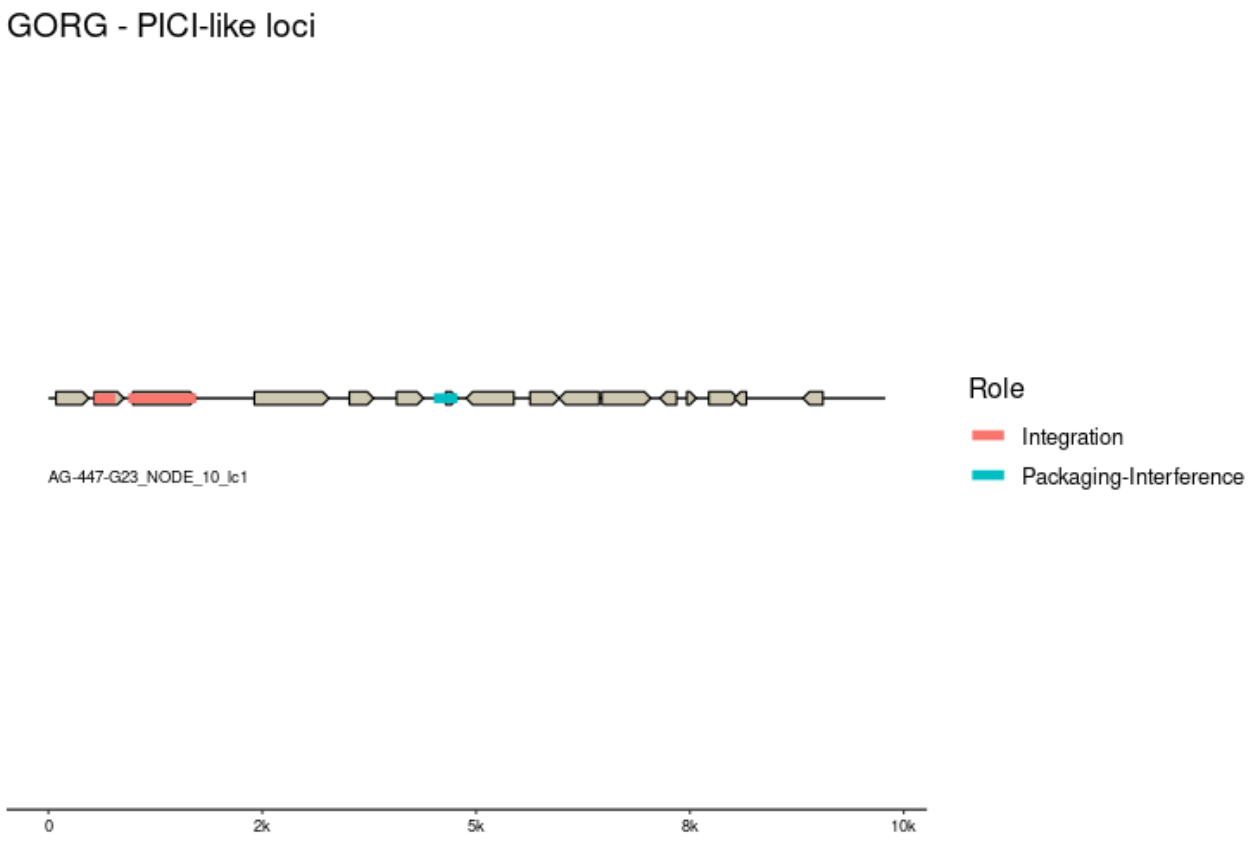


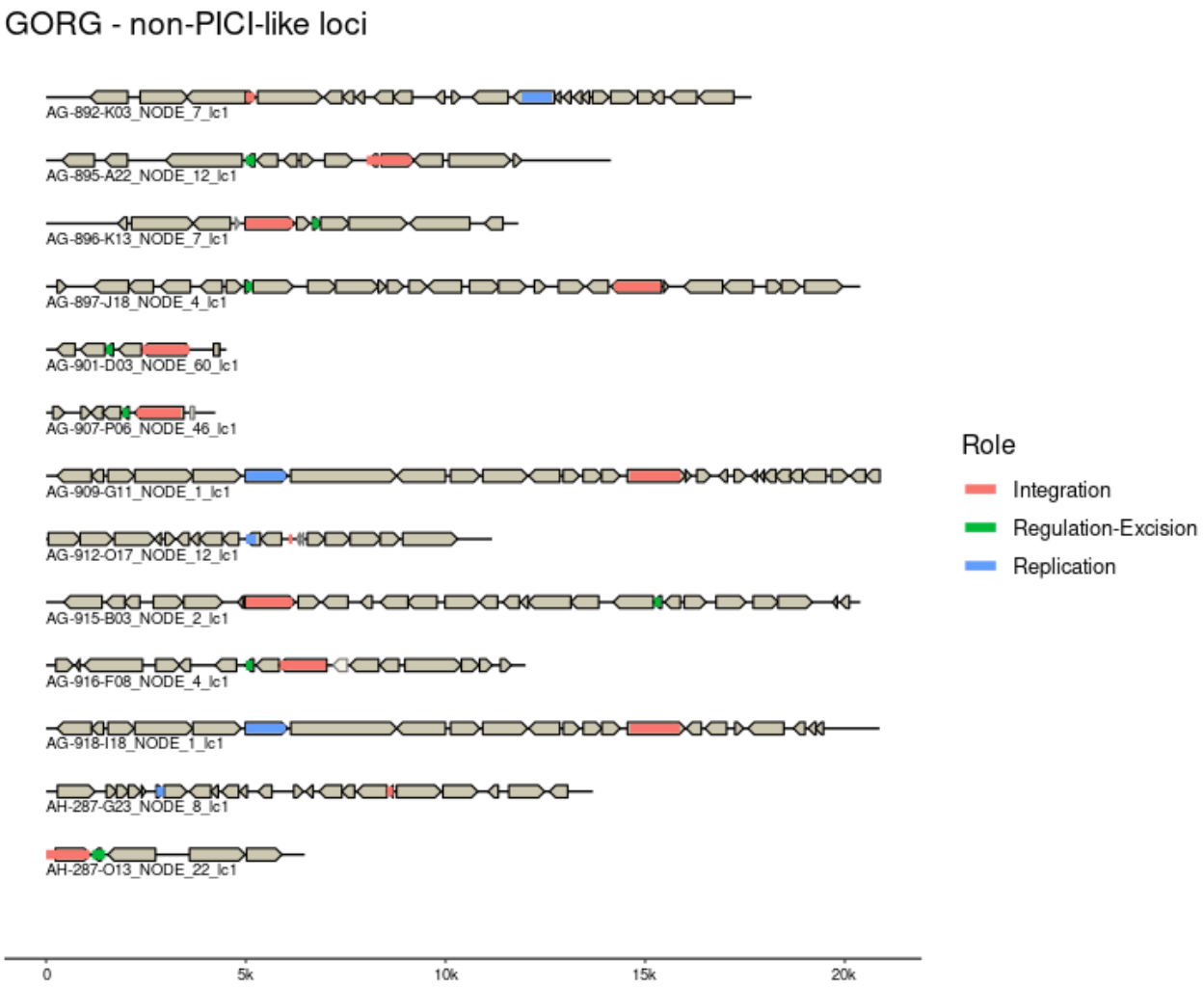

Tara Oceans - non-PICI-like loci
NYVG01000037.1_lc1
NYVG01000045.1_lc1
NYVN01000082.1_lc1
NYWC01000085.1_lc1
NYWP01000059.1_lc1
NYXE01000014.1_lc1
NYXN01000042.1_lc1
NYYX01000043.1_lc1
NZBW01000008.1_lc1
NZCM01000196.1_lc1
NZEW01000014.1_lc1
NZFK01000011.1_lc1
NZHS01000070.1_lc1
NZIO01000043.1_lc1
NZKN01000041.1_lc1
0
10k
20k
30k
Role
Integration
Regulation-Excision
Replication

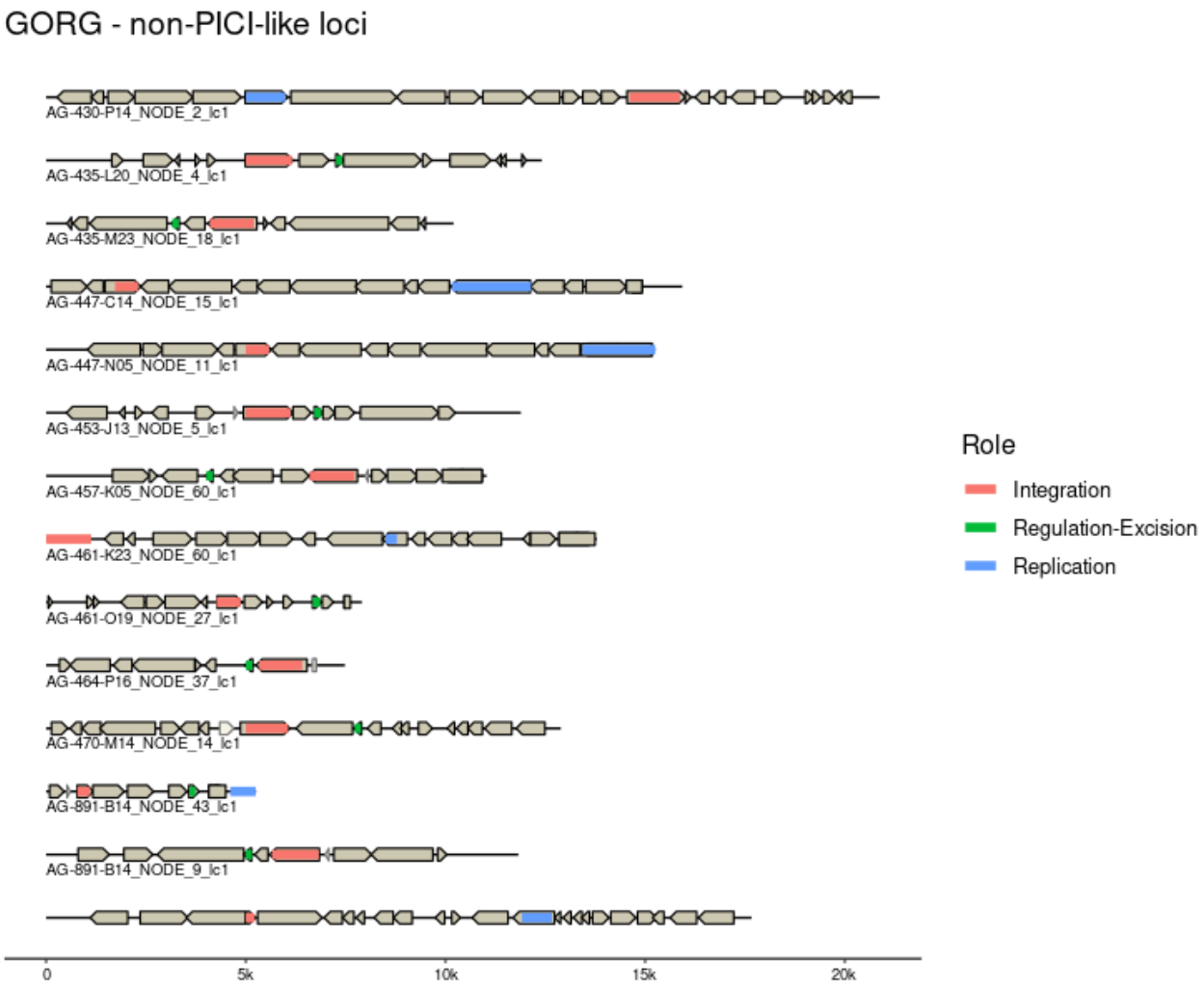
Tara Oceans - non-PICI-like loci



6.2 Appendix 2 - GORG




GORG - non-PICI-like loci
AG-892-K03_NODE_7_lc1
AG-895-A22_NODE_12_lc1
AG-896-K13_NODE_7_lc1
AG-897-J18_NODE_4_lc1
AG-901-D03_NODE_60_lc1
AG-907-P06_NODE_46_lc1
AG-909-G11_NODE_1_lc1
AG-912-O17_NODE_12_lc1
AG-915-B03_NODE_2_lc1
AG-916-F08_NODE_4_lc1
AG-918-I18_NODE_1_lc1
AH-287-G23_NODE_8_lc1
AH-287-O13_NODE_22_lc1
0
5k
10k
15k
20k
Role
Integration
Regulation-Excision
Replication


GORG - non-PICI-like loci
AG-430-P14_NODE_2_lc1
AG-435-L20_NODE_4_lc1
AG-435-M23_NODE_18_lc1
AG-447-C14_NODE_15_lc1
AG-447-N05_NODE_11_lc1
AG-453-J13_NODE_5_lc1
AG-457-K05_NODE_60_lc1
AG-461-K23_NODE_60_lc1
AG-461-O19_NODE_27_lc1
AG-464-P16_NODE_37_lc1
AG-470-M14_NODE_14_lc1
AG-891-B14_NODE_43_lc1
AG-891-B14_NODE_9_lc1
0
5k
10k
15k
20k
Role
Integration
Regulation-Excision
Replication

GORG - non-PICI-like loci