University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2018-11-29

Bioinformatic and phylogenetic analyses of retroelements in bacteria

Wu, Li

Wu, L. (2018). Bioinformatic and phylogenetic analyses of retroelements in bacteria (Doctoral thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/34667 http://hdl.handle.net/1880/109215

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Bioinformatic and phylogenetic analyses of retroelements in bacteria

by

Li Wu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOLOGICAL SCIENCES

CALGARY, ALBERTA

NOVEMBER, 2018

© Li Wu 2018

Abstract

Retroelements are mobile elements that are capable of transposing into new loci within genomes via an RNA intermediate. Various types of retroelements have been identified from both eukaryotic and prokaryotic organisms. This dissertation includes four individual projects that focus on using bioinformatic tools to analyse retroelements in bacteria, especially group II introns and diversity-generating retroelements (DGRs). The introductory Chapter I gives an overview of several newly identified retroelements in eukaryotes and prokaryotes. In Chapter II, a general search for bacterial RTs from the GenBank DNA sequenced database was performed using automated methods. It not only enlarged the collection of bacterial reverse transcriptases (RTs), but also revealed several new classes of RTs. In Chapter III, another automated search was performed to identify group II introns. All predicted introns were automatically folded and then manually refined. Other information, such as multiple copies and non-standard intron organisations, were also identified. Next, all introns were subjected in several analyses in order to depict common properties for each class, such as preferences in target sites and DNA strands. Using this enlarged dataset, Chapter IV aimed to resolve the phylogeny of group II introns and investigate whether the intron-encoded protein (IEP) and RNA portions coevolved. Among trees constructed from various datasets, such as using different sequence masks, smaller sampled subsets or morphological features, the hypothesis that the IEP and RNA coevolved was supported by comparisons among most trees, even though it seemed to be rejected by formal topology tests. Finally, Chapter V compiled and systematically classified the most recent set of DGRs, which can be used as a reliable reference to direct future DGR-related studies and experimental designs.

Preface

Chapter II of this thesis has been published as Zimmerly S and Wu L. "An Unexplored Diversity of Reverse Transcriptases in Bacteria". *Microbiol Spectr*. 2015;**3**(2):MDNA3-0058-2014.

(As stated in "ASM Journals Statement of Authors' Rights", an ASM author retains the right to reuse the full article in his/her dissertation or thesis.)

Chapter V of this thesis has been published as Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL, Du A, Guo H, Ghosh P, Miller JF, Zimmerly S. "Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey". *Nucleic Acids Res*. 2018;**46**(1):11-24.

(This article is available under the Creative Commons CC-BY-NC license and permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.)

Table of Contents

List of Tables

List of Figures

List of Abbreviations

aa	amino acid
Abi	abortive bacteriophage infection
AIC	Akaike information criterion
AIDS	acquired immunodeficiency syndrome
ASDSF	average standard deviation of split frequencies
AVD	accessory variability determinant
BIC	Bayesian information criterion
BLAST	basic local alignment search tool
bp	basepair
CaMV	cauliflower mosaic virus
cas	CRISPR-associated
cccDNA	covalently closed circular DNA
cDNA	complementary DNA
CH	conserved hypothetical
CL	chloroplast-like
CLec	C-type lectin
CPR	candidate phylum radiation
CRISPR	clustered regularly interspaced short palindromic repeat
CTE	C-Terminal extension
DGR	diversity-generating retroelement
DIRS	Dictyostelium intermediate repeat sequence
DNA	deoxyribonucleic acid
DPANN	Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea
DUF	domain of unknown function
E value	expect value
EBS	exon-binding site
EN	endonuclease
FGE	formylglycine-generating enzyme
G2L	group II intron-like
GTR	generalised time reversible
HBV	Hepatitis B virus
HIV	human immunodeficiency virus
HMM	hidden Markov model
HRDC	helicase and RnaseD C-terminal

ICR	internal complementary repeats
IEP	intron-encoded protein
Ig	immunoglobulin
IMH	initiation of mutagenic homing
IN	integrase
kb	kilo-basepair
ldtA	legionella determinant target A
LINE	long interspersed nuclear element
LTR	long terminal repeat
ML	maximum likelihood; mitochondrial-like
MP	maximum parsimony
mRNA	messenger RNA
msDNA	multicopy single-stranded DNA
MSL	MutS like
NCBI	National Center for Biotechnology Information
NJ	neighbour-joining
NR	non-redundant protein database (of GenBank)
nt	nucleotide
NT	non-redundant nucleotide database (of GenBank)
ORF	open reading frame
PAG	potential accessory gene
PBS	primer binding site
pgRNA	pregenomic RNA
PLE	Penelope-like element
PR	protease
PSI-BLAST	position-specific iterated basic local alignment search tool
PSSM	position specific scoring matrix
rcDNA	relaxed circular DNA
RdRP	RNA-directed RNA polymerase
RH	RNase H
RNA	ribonucleic acid
RNP	ribonucleoprotein
rRNA	ribosomal ribonucleic acid
RT	reverse transcriptase
SINE	short interspersed nuclear element
TE	transposable element

TEN	telomerase "essential" N-terminal
TER	telomerase RNA
TERT	telomerase reverse transcriptase
TG	target gene
TP	terminal protein
TPRT	target-primed reverse transcription
TR	template repeat
TRBD	telomerase RNA binding domain
tRNA	transfer RNA
TvpA	Treponema variable protein A
UPGMA	unweighted pair group method with arithmetic mean
UVR	unknown VR
VR	variable region
WGS	whole genome shotgun

Chapter I. Introduction

The diversity of reverse transcriptases

In the 1950s, the central dogma of molecular biology was formulated to explain the flow of genetic information in life forms: information is stored in DNA, passes from DNA to RNA through transcription, and then to protein through translation [1]. However, this was revised after the discovery of **r**everse **t**ranscriptase (RT) enzymes in RNA tumour viruses in 1970 [2, 3], which showed that information can also be transferred in the reverse direction from RNA back to DNA.

Since then, additional RT enzymes have been discovered in a wide range of organisms, although mainly in eukaryotes. In eukaryotes, a large number of RTs are associated with class I **t**ransposable **e**lements (TEs), which are thought to make up at least one third of mammalian genomes, and two thirds of the human genome [4, 5]. Class I TEs are also called retrotransposons, and they involve an RNA intermediate during transposition [6, 7, 8, 9]: the DNA is first transcribed to the RNA intermediate, and the RNA is reverse transcribed to produce a copy of the DNA, which is then inserted into other locations within the same genome. The best-known class I TEs include **l**ong **t**erminal **r**epeat (LTR) elements and non-LTR elements. Other types of class I TEs associated with an RT include ***P****enelope*-**l**ike **e**lements (PLEs) and ***D****ictyostelium* **i**ntermediate **r**epeat **s**equences (DIRS's) [6, 7, 8, 9, 10, 11]. Besides class I TEs, two groups of viruses, retroviruses and pararetroviruses, also encode a polymerase with RT activity [12]. Instead of remaining within the same genome before and after transposition like class I TEs, replicated viruses of these two groups leave the host cell and infect other cells. In contrast to both class I TEs and infective RT-encoding viruses that usually have multiple copies in the genome, another two types of RT-containing elements, **te**lomerase **r**everse **t**ranscriptase (TERT) and the *rvt* gene, exist as single-copy RTs, and they have either known or potentially useful cellular functions [8, 13, 14].

Prokaryotic RTs were discovered later than eukaryotic RTs. The first type of prokaryotic RT-containing elements identified were retrons, although their biological functions remained ambiguous until about 30 years after their initial discovery [15, 16, 17]. Of all prokaryotic RTs presently known, group II intron RTs are the only ones known to be retromobile and are the best understood [18]. There are also other types of prokaryotic RTs with less clear or unknown functions, such as group II like RTs, **d**iversity-**g**enerating

retroelements (DGR), **a**bortive **b**acteriophage **i**nfection (Abi) related RTs, and many unclassified groups [8, 19, 20, 21, 22].

The large abundance and diversity of both eukaryotic and prokaryotic RTs have attracted much attention, not only because some RTs can increase genomic diversity by replicating and spreading throughout the genome (e.g. retrotransposons, group II introns), but also because some have essential functions to the host (e.g. TERT) or could interrupt normal gene expression and cause diseases (e.g. SINE, see below). The general topic of this thesis dissertation involves the properties and evolution of bacterial RTs. To give a context for projects included in this dissertation, this chapter will introduce many major types of RTs in both eukaryotes and prokaryotes, including their structures, properties and evolutionary relationships.

Eukaryotic RTs

Retroviruses and LTR retrotransposons

RTs were first discovered from retroviruses [2, 3], which are a family of viruses that have an RNA genome and infect cells through a DNA intermediate. The most studied virus of this family is the **h**uman **i**mmunodeficiency **v**irus (HIV), which causes **a**cquired **i**mmuno**d**eficiency **s**yndrome (AIDS). The life cycle of retroviruses generally resembles that of other viruses, including virus attachment, entrance of the viral genome into cells, expression of viral genes, assembly and release of viral particles [12, 23, 24]. Unlike other viruses with a DNA genome, the RNA genome of retroviruses is first reverse transcribed into DNA, which is then integrated into the host genome to form a provirus, and finally followed by the expression of viral genes from the proviral DNA [12, 24].

The genome of retroviruses commonly contains three genes, named *gag, pol* and *env* respectively. The *gag* gene codes structural core capsid proteins, the *pol* gene encodes an enzyme with RT, RNase H (RH) and integrase (IN) activities, and the *env* gene encodes the envelope protein [6, 12, 25]. The *gag* and *pol* genes are fused, and either of them could also encode a protease (PR) in order to process their primary protein product [25]. Terminal direct repeats are present at both ends of the viral genome, surrounding the coding region [12, 25].

Long **t**erminal **r**epeat (LTR) retrotransposons are a type of class I TEs, and they resemble the gene structure of retroviruses. They were named after the presence of a

direct repeat on both sides of a coding region, which codes a *gag*, *pol* and sometimes an *env* gene that are equivalent to those of a retrovirus (Figure 1A) [7, 25]. They are commonly 5-7 kb in length while the repeated sequences are often a few hundred base pairs long [7, 25]. LTR retrotransposons are commonly found in animals, fungi, protists and plants in high copy numbers, and they can be divided into subclasses based on sequence similarity and gene order [7].

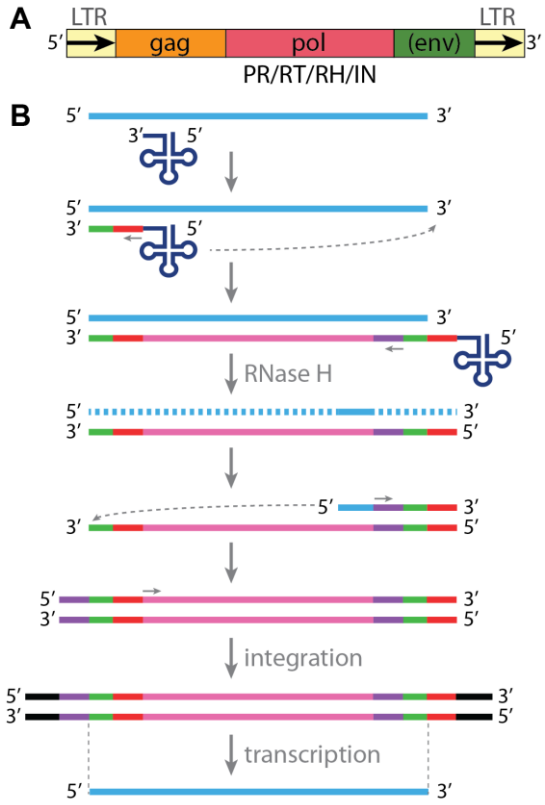


Figure 1. Structure and mechanism of LTR retrotransposons.

A) LTR retrotransposons contain repeated sequences at both termini flanking a coding region, which consists of a *gag*, *pol* and sometimes *env* genes, while the protein product of the pol gene has PR, RT, RH and IN activities. **B)** Transposition of LTR elements begins with its RNA transcript (blue). A tRNA of the host binds to its 5' region and primes cDNA synthesis of a short piece that includes unique sequence at the 5' end (red) and a repeated sequence (green). It then moves to the 3' end of the template RNA through base pairing of the short repeat (green), and cDNA synthesis continues to produce the full-length cDNA (pink), and purple indicates a short unique sequence at the 3' end of template. RH then degrades the RNA template, leaving only a short tract that primes DNA synthesis of the 3' portion of the other strand (purple+green+red), which later moves to pair with the 3' portion of the first strand and finishes DNA synthesis of the entire second strand. Finally, the double stranded DNA is integrated into chromosomal DNA (black), and future RNA transcription (blue) starts and ends within the LTR sequences. Small grey arrows indicate the orientation of DNA synthesis. Figure adapted from reference [25].

Both retroviruses and LTR retrotransposons contain a tRNA primer binding site (PBS) downstream of the 5' repeat (Figure 1B) [7, 26]. A host tRNA first binds to the PBS and serves as a primer for reverse transcription of the 5' portion, including a short repeat that can pair with the 3' end and allow cDNA synthesis of the entire transcript. The RNA portion is degraded by RH activity excluding a poly-purine tract that has RH resistance. This tract then primes cDNA synthesis of the second strand along the 5' portion of the template including the short repeat again. The newly synthesised short DNA base pairs to the short repeat of the 3' side of the template, followed by cDNA synthesis of the entire second strand. In the end, the double-stranded DNA is inserted into host chromosome by the IN activity (Figure 1B).

Retroviruses and LTR elements have similarities in both sequence and mechanism, and they are closely related. The main difference is that copies of retroviruses can leave the current host and infect other cells, while copies of LTR retrotransposons are restricted to move within the same genome [25]. Therefore, LTR elements are thought to be inactivated provirus that lost the ability to exit the host [27]. However, LTR elements could still have infectious properties and be transferred horizontally if they could use a foreign envelope protein or become associated with other infectious agents [12, 25, 28, 29, 30].

Non-LTR retrotransposons – LINE and SINE elements

The second major group of eukaryotic RTs are non-LTR retrotransposons, which also belong to class I TEs and contain diverse groups. They do not contain terminal repeats that are present in LTRs, but instead there is an adenosine-rich region (usually a poly-A tail) at the 3' end (Figure 2A) [6]. The two most major classes of non-LTR retrotransposons are **l**ong and **s**hort **i**nterspersed **n**uclear **e**lements (LINEs and SINEs), which are respectively several kilobases and a few hundred bases in length. Unlike LTRs, there is no viral form of non-LTR elements. Based on sequence, non-LTR elements are the most closely related to group II introns (see below) and both encode a protein with RT and endonuclease (EN) activities. Also because both use a target primed reverse transcription (TPRT) mechanism, non-LTR elements are thought to have originated from group II introns [6, 31, 32, 33].

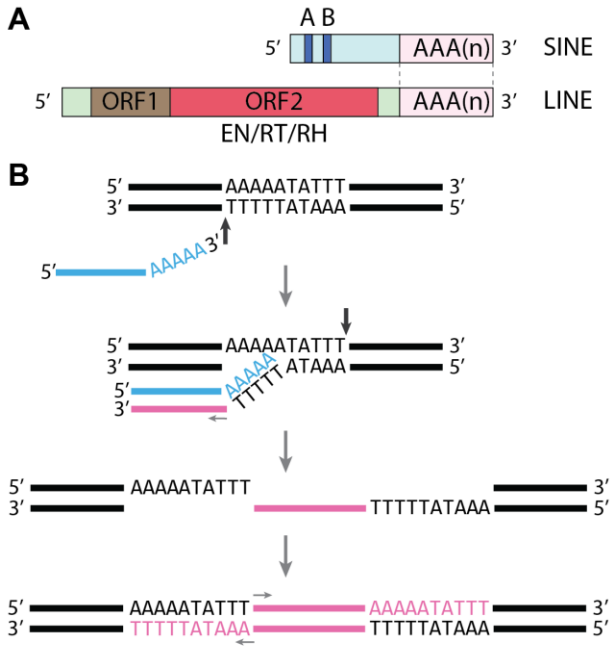


Figure 2. Structure and mechanism of non-LTR retrotransposons.

A) LINE and SINE elements are major types of non-LTR retrotransposons, and both have an A-rich region at the 3' end, often being a poly-A tail. SINE elements do not encode any protein. They contain the RNA polymerase III promotor Boxes A and B at the 5' end, and usually share some sequence similarity at the 3' end to LINE elements. LINE elements usually encode two ORFs, while ORF2 produces a protein with EN, RT and RH activities. **B)** Non-LTR retrotransposons tend to target A/T-rich regions of chromosomal DNA. The bottom strand at the target is cleaved by EN, and through base pairing of the 3' end of the non-LTR element, the first strand of cDNA (pink) is synthesised. The other strand of the target DNA is then cleaved, allowing DNA synthesis of the second strand. Small grey arrows indicate the orientation of DNA synthesis. Figure adapted from reference [25].

LINE elements are widespread in eukaryotes and can be divided into subtypes based on structural features and RT-based phylogenies [34, 35, 36]. However, most copies of LINE elements lack sequences from the 5' end in various lengths and thus appear to be incomplete and lose the ability to transpose [25, 37]. In humans, the only active class of LINE element is called LINE1 or L1, which make up about one fifth of the human genome [38, 39].

A typical LINE element contains two **o**pen **r**eading **f**rames (ORFs) named ORF1 and ORF2 (Figure 2A). ORF1 codes for an RNA-binding protein, ORF2 codes for a protein with EN, RT and sometimes RH activities [25, 40]. Through the assistance of the protein translated from ORF2, the RNA transcript is first associated with chromosomal DNA at an A/T-rich region during transposition (Figure 2B). The nuclease activity creates a

break and releases a short string of T from the chromosomal DNA, which base pairs to the poly-A tail of the RNA intermediate and primes reverse transcription of the LINE element cDNA. The nuclease then cleaves the second strand of the chromosomal DNA a few nucleotides away from the first break, while the RNA is degraded by RH activity followed by DNA synthesis of the second strand. Finally, the double stranded LINE DNA fully integrates into chromosomal DNA through the host repair system [25] (Figure 2B). Although most LINE elements are not site-specific, there are exceptions including subtypes CRE, R2 and R4 that are specific for rRNA genes or simple repeats [33, 41, 42, 43, 44, 45].

SINE elements are shorter than LINE elements and do not encode any ORF (Figure 2A). Their transposition requires the protein encoded by LINE elements [25, 46]. SINEs are origin-specific and were derived from small RNA genes such as tRNA, 7SL RNA and 5S RNA, which are all recognised by RNA polymerase III [7, 46]. The 5' region of SINE elements contains the promotor boxes A and B of RNA polymerase III, the internal regions of SINEs vary depending on the origin, and the 3' region is usually related to the 3' end of LINE elements to ensure the recognition by LINE-related proteins, which all potentially imply coevolution between LINEs and SINEs (Figure 2A) [25, 46, 47, 48]. In humans, the best-known SINE retrotransposon is the Alu element, which is the most active and abundant transposable element that exists with over one million copies [49, 50]. Disruptive insertions of Alu elements are associated with many human diseases, such as hemophilia and breast cancer [51, 52, 53, 54, 55, 56].

PLEs

P*enelope*-**l**ike **e**lements (PLE) are another unusual group of class I TEs that have a single ORF coding for an RT and a GIY-YIG (or Uri) EN domains (Figure 3A), in which the latter serves as an integrase and is also found in bacterial group I introns and UvrC bacterial DNA-repair endonucleases [57, 58, 59, 60, 61]. Their RTs are less similar to other retrotransposon RTs but have been shown to be more closely related to TERT (see later) [58, 60, 62, 63]. PLEs have many variations in their structure, some lack the EN domain, some have either direct or inverted repeated sequences at both termini, and some were even found to contain introns [7, 31, 64, 65]. PLEs that encode an EN domain insert randomly in the genome but have a preference for AT-rich regions, resembling non-LTR elements [6, 59]; PLEs that do not encode an EN domain prefer to

insert into telomeric regions of chromosomes and are hypothesized to be associated with TERT (Figure 3B) [65]. PLEs have been found in many eukaryotic species, but their distribution is rather dispersed, which indicates a high degree of lineage losses [31, 59, 61, 65].

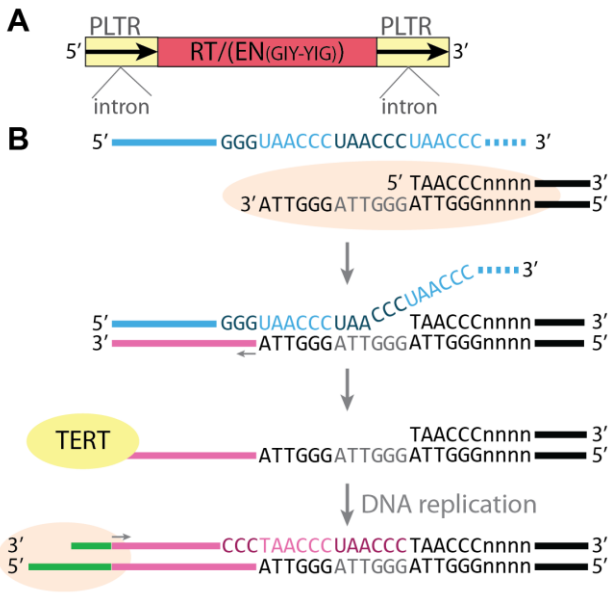


Figure 3. Structure of PLEs and an EN-independent model of retrotransposition.
A) A schematic representation of PLEs that vary in gene organisation and order. This example shows a PLE with direct repeats (PLTR) at both termini resembling those of an LTR element but may be interrupted with introns. The PLE ORF encodes an RT and sometimes a GIY-YIG EN domains. **B)** Model for retrotransposition in EN-independent PLEs associated with TERT. The PLE RNA (blue) anneals to single stranded telomeric repeats, and the priming process begins at a G-rich region to synthesise the first strand of PLE DNA (pink). TERT is then involved to generate new telomeric repeats (green), and the second strand of PLE DNA is synthesised through regular DNA replication at the same time. Pale orange oval, proteins that normally cap at telomeres. Text in different shades highlights each telomeric repeat unit. Small grey arrows indicate the orientation of DNA synthesis. Figure adapted from references [57, 65].

DIRS

D*ictyostelium* **i**ntermediate **r**epeat **s**equences (DIRS) were first discovered in the slime mould *Dictyostelium discoideum* as an unusual type of RT-containing element [66]. They likely originated from LTR elements, but their RT sequences are more divergent compared to RTs from both LTR and non-LTR elements [64]. Unlike other retrotransposons, they do not encode a protein with IN activity, but instead encode a Gag protein, a protein with RT and RH activities, and a tyrosine recombinase that is

similar to that of some DNA transposons [66, 67]. They have repeated sequences in inverted orientations at both termini. Within the element, there is also a repeated segment that forms **i**nternal **c**omplementary **r**epeats (ICR), which serves as a template during reverse transcription and results a circular RNA intermediate (Figure 4A) [64, 66, 67]. The insertion of DIRS into the chromosome is through recombination rather than integration with the contribution of the tyrosine recombinase (Figure 4B) [64, 66, 67]. Many DIRS elements have been discovered in different organisms, and they vary in the locations of LTR, ICR as well as the arrangement of their encoded ORFs [64, 68].

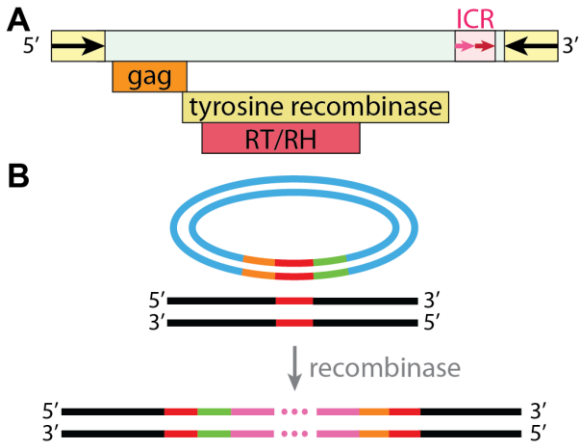


Figure 4. Structure and mechanism DIRS.

A) Gene structure of DIRS-1 from Dictyostelium discoideum. It contains reversed repeat sequences at both termini, flanking a coding region with three overlapping ORFs, which encode a Gag protein, a tyrosine recombinase, and a protein with RT and RH activities. At its 3' end, there is a repeated segment forming into an internal complementary repeats (ICR). **B)** The ICR is responsible for a circular RNA intermediate (blue), which is inserted into chromosomal DNA through the recombinase. Red, target site. Orange and green, regions flanking the target site on DIRS. Pink, retroelement in DNA. Figure adapted from reference [67].

Pararetroviruses – hepadna- and caulimoviruses

In addition to retroviruses that were described earlier, the group of pararetroviruses is another type of viruses that encode an RT. This group contains the hepadnaviruses family that infects animals and the caulimoviridae family that infect plants, which are represented by **H**epatitis **B** **v**irus (HBV) and **ca**uliflower **m**osaic **v**irus (CaMV) respectively [69, 70]. Although these two groups of viruses encode an RT and use reverse transcription during replication, they are different from retroviruses as they have a DNA genome rather than an RNA genome, and do not integrate into host

chromosomes [12, 69, 70, 71, 72].

The genome of hepadnaviruses is a partially double-stranded **r**elaxed **c**ircular DNA (rcDNA), which is converted into a **c**ovalently **c**losed **c**ircular DNA (cccDNA) when being transported into the nucleus [70]. The cccDNA serves as a template to produce viral RNAs including a pregenomic RNA (pgRNA), which is finally converted to viral rcDNA after virus replication [70, 73]. The representative HBV polymerase (P protein) contains multiple domains: terminal protein (TP), spacer, RT and RH [70]. Reverse transcription of the pgRNA is primed by an OH group from a specific tyrosine residue in the TP domain, and this step is unique in hepadnaviruses [70, 73, 74].

Similar to hepadnaviruses, caulimoviruses also have a double-stranded DNA genome in an open circular form and have interruptions on both DNA strands [70]. Inside the host cell nucleus, these interruptions are repaired by the host machinery and the repaired new sequence is then used as a template to express viral proteins and also to produce pgRNA [70, 72]. The polymerase of caulimoviruses contains PR, RT and RH domains [70]. All known caulimoviruses only uses a methionine initiator tRNA primer for reverse transcription of the pgRNA. This resembles retroviruses and retrotransposons, although they can use various tRNAs as the primer [70, 72].

TERTs

Different from all RT-containing elements introduced so far that are either selfish DNAs or parasitic viruses, **te**lomerase **r**everse **t**ranscriptase (TERT) has an essential role in living cells to maintain the telomere length. TERT is highly conserved and contains four domains: a **t**elomerase "**e**ssential" **N**-terminal (TEN) domain that interacts with both the single stranded telomeric DNA and TER at the same time [75, 76], a **t**elomerase **R**NA **b**inding **d**omain (TRBD) that contains motifs forming the RNA-binding pocket and also binds to TER [77], an RT domain that is involved in nucleotide addition [78, 79], and a **C**-**T**erminal **e**xtension (CTE) domain that stabilises the RNA-DNA duplex as well as the RT domain that binds to DNA (Figure 5A) [80]. The RT domain has a similar motif organisation to classic RT enzymes, and has been found to be closely related to RTs of other retrotransposons, especially PLEs [58, 63, 65, 81, 82, 83].

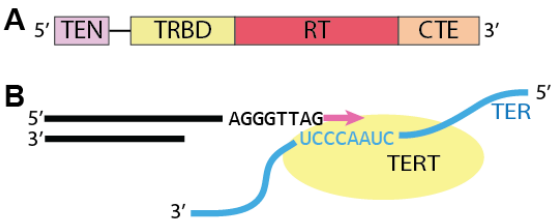


Figure 5. Structure of TERT and interaction between telomerase and telomere.
A) TERT contains four domains: Telomerase "essential" N-terminal (TEN) domain, telomerase RNA binding domain (TRBD), RT domain and C-terminal extension (CTE) domain. **B)** Simplified representation of telomerase and telomeric DNA, in which the RNA template from telomerase RNA (TER, blue) pairs to the single stranded DNA at telomere (black), which enables DNA synthesis of telomeric repeat (pink arrow). Other associated proteins during this process is not shown. Figure adapted from reference [78].

Both TERT and telomerase RNA (TER) are essential cores of a telomerase, which is a ribonucleoprotein (RNP) enzyme of the telomerase complex. This complex is responsible for catalysing the addition of repetitive nucleotides to the end of telomeres of a chromosome using an RNA template provided by TER to maintain the length of telomeres as well as to protect them from being mistaken as damaged DNA and degraded (Figure 5B) [84, 85, 86, 87, 88, 89, 90]. Mutations in either TERT or TER of telomerase have been shown to be associated with many human diseases [78, 91]. TERT is unique from most other RT-containing elements in eukaryotes, as it exists as a single copy and has crucial functions to the cell.

rvt elements

After TERT, reverse transcriptase-related (*rvt*) genes were discovered as the second type of single-copy RT and is predicted to have a cellular function [14]. They were named after the fact that they contain a *rvt* conserved domain [14]. Currently, almost all *rvt* genes are found in eukaryotes including protists, fungi, plants and animals, and only limited examples have been found in bacterial genomes. The extreme difference in the abundance of *rvt* genes implied that the bacterial copies more likely originated from a rare horizontal transfer from eukaryotes to prokaryotes [14].

A typical *rvt* is ~1 kbp in length and contains an RT domain, which is conserved in sequence among different individuals, but shows little similarity to other RTs. This RT domain is flanked by extensions at both N- and C-termini of ~300 and ~200 amino acids

respectively, but no known protein homologs could be identified except a coiled-coil motif at the N terminus [14]. A more recent *in vitro* study has shown this coiled-coil motif is responsible for forming multimers in solution [92]. The C-terminal region could be associated to protein priming [92], as purified proteins showed template-independent terminal transferase activity and can polymerise both NTPs and dNTPs to the 3'-OH of a given nucleotide primer, while NTPs are preferred, which puts *rvt*s more closely related to TERT [14, 92]. The *rvt* protein does not contain an EN domain or other equivalents, agreeing with their single-copy property. Nonetheless, the biological function of *rvt* genes currently remains unclear, although it was proposed to be related to stress response or other non-essential cellular processes [14].

Prokaryotic RTs

Group II introns

Many types of RTs have been identified in prokaryotes. At present, the best-known type of prokaryotic RTs is from group II introns. Group II introns are retroelements with catalytic activities consisting of an RNA ribozyme and an **i**ntron-**e**ncoded **p**rotein (IEP) (Figure 6) [18, 93, 94]. They are widespread in bacteria and archaebacteria, and have also been found in organellar genomes of protists, fungi and plants [95, 96, 97, 98, 99, 100, 101]. The RNA ribozyme is usually 500-800 bases in length. Although they do not share a high degree of sequence similarity, they generally form six conserved domains [102, 103, 104]. In contrast, the IEP is usually 1-1.5 kb in length. A typical IEP consists of multiple domains while some have sequence similarity: a RT domain that can be aligned with other IEPs as well as other types of RTs; a maturase domain (X) that resembles a thumb domain of a polymerase, but without sequence similarity to thumb domains in other proteins; a DNA-binding domain (D) and sometimes an EN domain (Figure 6A) [18, 105, 106, 107, 108, 109, 110, 111].

Group II introns can splice out from flanking sequences and have the ability to insert into specific locations in the genome, which is known as retrohoming. Although the RNA alone is capable of self-splicing *in vitro* under specific conditions, the IEP is essential for both processes *in vivo*. While the RT and X domains are required for splicing, all domains are required for retrohoming [110, 111, 112, 113, 114, 115, 116].

The splicing reaction of group II introns involves two transesterification steps (Figure 6B)

[117, 118, 119, 120]. In the first step, the 2'-OH of the bulged A of the intron RNA attacks and cleaves the 5' exon-intron junction, and forms a lariat intermediate with the intron RNA and 3' exon. In the second step, the 3'-OH from the 5' exon attacks the 3' splicing site and releases the lariat intron and ligated exons (Figure 6B). Under *in vivo* conditions, the IEP assists the intron RNA to be properly folded and remains bound to the intron lariat after splicing [103, 110, 112, 114, 121, 122, 123].

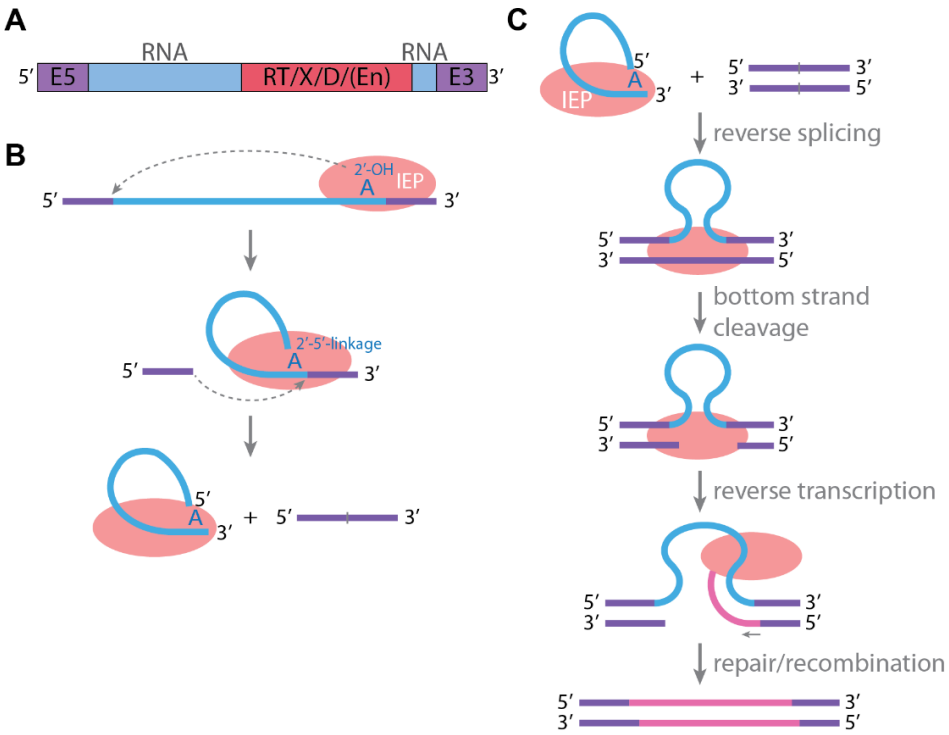


Figure 6. Structure and mechanisms of group II introns.

A) A typical group II intron consists of an RNA ribozyme (blue) and an intron-encoded protein (IEP, red), which contains RT, X, D and sometimes EN domains. Purple represents flanking 5'- and 3' exons (E5 and E3). **B)** Splicing of group II introns. First, the bulged A from the RNA (blue), which provides its 2'-OH and attacks the 5' exon (purple), forming an intron lariat connected by 2'-5'-linkage. Next, the 5' exon attacks the 3' exon, resulting in ligated exons (purple) and the intron lariat is bound to the IEP. **C)** Retrohoming of group II introns. The intron-IEP complex moves to the insertion site of genomic DNA (purple), and the intron RNA first reverse-splices into the top strand. The IEP then cleaves the bottom strand, allowing DNA synthesis of the intron. Finally, double stranded intron DNA is integrated through host repair and recombination systems. Small grey arrow indicates the orientation of DNA synthesis. Figure adapted from reference [22].

The RNA-protein (RNP) complex after splicing is capable of the retrohoming reaction, in which the RNA sequence is inserted into other loci in the genome using the target primed reverse transcription (TPRT) mechanism (Figure 6C) [93, 108, 124, 125, 126]. First, the complex binds to the insertion site of the double stranded target DNA and the intron RNA reverse splices into the sense strand of the target [127]. The EN domain of the IEP then cleaves the other strand downstream of the insertion site, providing a primer for the RT to synthesise DNA from the intron RNA. Finally, the intron RNA is degraded and sense DNA is synthesised by the host repair system (Figure 6C) [108, 128, 129, 130].

In some introns that lack the EN domain, an alternative mechanism is required for priming the reverse transcription. Such introns usually invade the DNA strand that serves as a template for the lagging strand during DNA replication, and the priming process thus uses either primase or Okazaki fragments on the lagging strand for reverse transcription [131, 132, 133, 134, 135, 136]. Occasionally through other mobility mechanisms, they can also insert into ectopic sites that do not have much sequence similarity to normal homing sites [137, 138, 139, 140].

Group II like RTs and CRISPR/Cas association

Group II like (G2L) RTs are highly similar to typical group II RTs in sequence. Unlike group II introns, G2Ls do not contain an RNA ribozyme at either side of the RT gene, and it is yet unclear whether they are associated to a different type of ribozyme. Currently, there are five G2L classes, G2L1-5, based on the RT phylogeny [20]. G2Ls generally contain motif 0 of the RT domain and the X domain, which are often not present in other types of prokaryotic RTs (Figure 13 of Chapter II). Many G2L classes contain extensions of several hundreds of amino acids on one or both sides of the RT, but whether these extensions have functional motifs remains unknown (Figure 13 of Chapter II) [20].

Both G2L1 and G2L2 RTs were found to be associated with CRISPR (**c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats) elements, as they are either fused or adjacent to a *cas1* (**C**RISPR-**as**sociated) gene [20]. The CRISPR/Cas system is known as the prokaryotic adaptive immune system [141, 142]. DNA from infective phages will be integrated into the CRISPR array during infection, ensuring recognition and

inactivation of future phages [143, 144, 145]. The Cas1 protein has nuclease activity and forms a complex together with the Cas2 protein, and the complex is essential during the integration of new spacers into CRISPR arrays [20, 142, 144, 145, 146, 147, 148, 149]. G2L1 and 2 are thus thought to be involved in certain types of the CRISPR/Cas systems and may use a TPRT-like mechanism to insert phage DNA into CRISPR arrays [20, 22, 146, 150, 151]. However, properties of G2L3-5 remain unclear.

Retrons

The class of retrons was the first identified type of RT-containing elements in bacteria [15, 152, 153]. A retron is usually 2 kb in length and contains three genes, *ret, msd* and *msr* (Figure 7) [15, 152, 153, 154]. The *ret* gene encodes a retron-specific RT, while the *msd* and *msr* genes together form a unique single-stranded DNA/RNA hybrid, termed **m**ulticopy **s**ingle-stranded DNA (msDNA), which has been reported to accumulate in the host with up to thousands of copies (Figure 7) [155, 156]. In the host, the RT binds to the RNA transcript of *msd* and *msr*, and uses a 2'-OH group provided by a G residue from *msr* as a primer to reverse transcribe a segment of this RNA transcript (Figure 7). The portion of RNA that served as a template (Figure 7, green) for reverse transcribing the DNA is later degraded by RH activity from the host, leaving a chimeric DNA-RNA complex (Figure 7) [156, 157, 158].

The natural function of retrons has remained unknown for about 30 years since its discovery. Retrons have been found to be inherited both vertically and horizontally, implying the function different retrons may vary depending on the specific environment of each host [15, 154]. A more recent study showed that the absence of msDNA in the pathogen *Salmonella* caused a dysregulation of proteins involved in the anaerobic metabolism that is required by this organism, and provided some evidence that retrons could serve as regulatory factors for protein abundance [159].

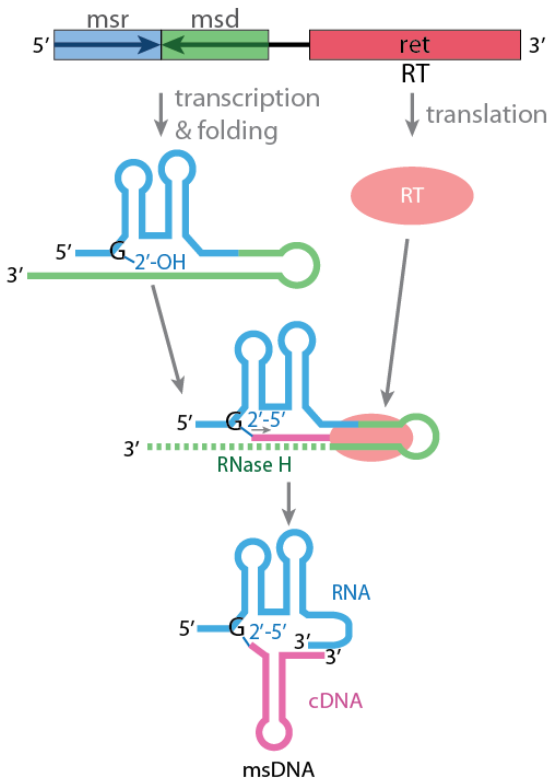


Figure 7. Gene structure and msDNA formation of retron.

A retron contains a ret gene that encodes a retron-specific RT. Upstream of the *ret* gene, there are other two genes, *msr* and *msd*. The portion of RNA transcript containing the *msr* (blue) and *msd* (green) folds into a particular secondary structure, allowing the branching guanosine residue (G) to provide its 2'-OH to prime the reverse transcription of the *msd* gene, while RH degrades the RNA template at the same time. Reverse transcription stops at a fixed point, resulting in a msDNA (multicopy single-stranded DNA) in which the cDNA (pink) is linked to the RNA (blue) by 2'-5' linkage. Small grey arrow indicates the orientation of DNA synthesis. Figure adapted from reference [156].

DGRs

Diversity-**g**enerating **r**etroelements (DGRs) are retroelements that consist of multiple gene components in addition to the RT. They have been discovered in bacteria, phages and plasmids [19, 62, 160, 161]. The best experimentally studied DGR is from the *Bordetella* BPP-1 phage and is often used as a model DGR [62]. This DGR contains an RT, a **t**arget **g**ene (TG), an **a**ccessory **v**ariability **d**eterminant (*avd*) gene, as well as two sequence repeats. One is the **t**emplate **r**epeat (TR) and the other is the **v**ariable **r**egion (VR) that is usually located at the 3' end of the TG and is similar in sequence with the TR (Figure 8).

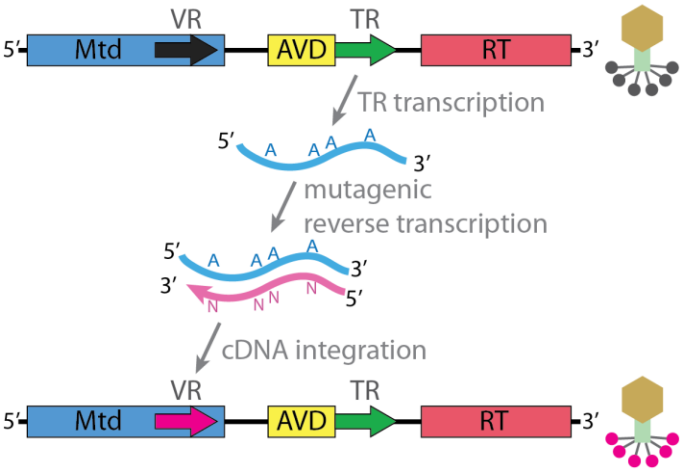


Figure 8. Architecture and mutagenic retrohoming of the *Bordetella* BPP-1 phage DGR. This DGR consists of multiple components in the order of a TG that contains the VR, an AVD, a TR and an RT. The TG codes for a Mtd protein and locates at tips of phage tail fibres and is responsible for cell binding during infection. During mutagenic retrohoming, the RT reverse transcribes the RNA of TR (blue) while all A's in the template RNA are randomly mutated to any nucleotide. The cDNA replaces the old VR, resulting in the sequence at the C-terminus of TG being diversified. Correspondingly, the DGR contributes in increasing phage adaptation by generating new variants of the Mtd protein. Figure adapted from reference [161].

DGRs are immobile but have beneficial functions to the host through mutagenic retrohoming [62, 162, 163, 164]. First, the RT reverse transcribes the RNA of TR, while every A of the template is randomly mutated to any nucleotide. The newly synthesised DNA thus has a different sequence with random A-to-N mutations compared to the original TR, and is then integrated into the TG by replacing the old VR. As a result, the new TG will have a different DNA sequence at the 3' end, and accordingly the amino acid sequence at the C terminus of TG will be also different (Figure 8).

In the case of the *Bordetella* phage DGR, the TG codes for the Mtd protein, which locates at the tip of phage tail fibres and can bind to the bacterial receptors on the cell surface during phage infection [62, 163, 165, 166, 167, 168]. Therefore, the *Bordetella* phage DGR benefits the host phage adaptation by generating variants of the Mtd protein (Figure 8). Studies of other DGRs have also indicated that the TG is often associated to cell recognition or binding functions, and DGRs therefore are generally beneficial to the host by diversifying the TG and increasing the host adaption [62, 169, 170].

Abi-related RTs

Some bacteria have phage immunity mediated by the **a**bortive **b**acteriophage **i**nfection (Abi) system, in which infected cells interrupt phage development, resulting in little or no phage progeny and the death of infected cells [171]. Some identified Abi systems have been reported to encode proteins containing an RT domain , including AbiA and AbiK and an Abi analogue named Abi-P2 [172, 173, 174]. AbiA and AbiK proteins share some sequence identity, and both systems are thought to function similarly by preventing phage replication [20, 22, 171, 175, 176, 177, 178].

The AbiK protein encodes the RT domain in the N-terminal region, and its C-terminal region is critical to normal AbiK activity and is thought to contain the thumb domain [179]. The AbiK protein has been shown to have polymerase activity *in vitro*. But instead of normal reverse transcription processes, AbiK uses a tyrosine residue from itself as a primer to synthesise a random DNA sequence with neither a DNA nor an RNA template [180]. Although both AbiA and Abi-P2 proteins have the RT domain located at the N-terminal region while their C-termini remain uncharacterised, neither has been shown to have a similar activity to that of AbiK.

Uncharacterised RTs

In addition to RTs mentioned above, there are in fact many more uncharacterised RTs, with some of them falling into "unknown" groups [20, 21]. They are generally believed to be functional as they either have long extensions in addition to the RT domain that can potentially contain protein motifs or are fused to other known protein motifs [20]. There is currently no enough information to fully understand these unknown RTs, but it is expected that they will be better studied in the future when more examples are available.

Evolutionary relationship among RT-associated elements

Various RT-containing elements from both eukaryotes and prokaryotes are thought to have originated from the same ancestor because they have similar sequences and functions [8, 32, 81, 106, 147, 181, 182]. Prokaryotic RTs are often favoured to be the root for all RTs, not only because of their simpler form in which the RT is often the only domain, but also because they have a greater similarity in sequence to the **R**NA-**d**irected **R**NA **p**olymerase (RdRP) of RNA viruses, which is usually used as an outgroup (Figure

9) [32, 81]. Eukaryotic RTs are more complex but were likely derived from the same core structure consisting of *gag* and *pol* genes through gain or loss of functions (Figure 9) [32]. However, comprehensive phylogenetic studies across all RTs are challenging, as conventional sequence-based studies are largely limited by the lack of phylogenetic signals, and deep branches across diverse groups of RTs remain ambiguous [8, 20, 93, 147].

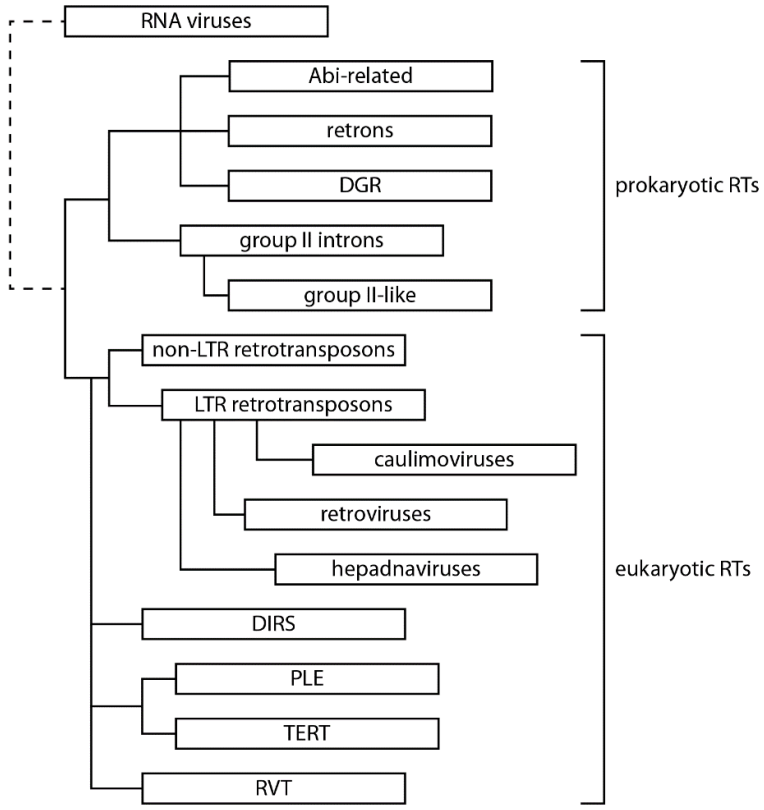


Figure 9. Relationships between diverse RT groups.

RTs are widespread in both eukaryotes and prokaryotes, and they were hypothesized to share the same common ancestor with RNA viruses. However, clear relationships between different RTs remain unresolved due to limited phylogenetic signal. Figure adapted from references [32, 81, 197] with modifications.

Among the diverse groups of RTs, group II introns are generally agreed to be the ancestor of non-LTR retrotransposons because of similarities in their protein sequences and both use a TPRT-based mechanism (Figure 10) [32, 33, 93, 100, 106, 183]. Group II introns have gained much attention because of their relationship to the eukaryotic

spliceosome as they both involve two transesterification reactions during splicing and have parallel structural arrangements (Figure 10) [100, 184, 183, 185, 186, 187, 188, 189, 190, 191]. First, an equivalent to the bulged A motif in DVI of group II introns can be observed in the spliceosome by pairing the **s**mall **n**uclear RNA (snRNA) U2 to the branch-point region of the intron; second, equivalents to the catalytic triad that forms the active site as well as the AC-bulge in the DV of group II introns [192, 193, 194, 195, 196] also exist in the snRNA U6 of the spliceosome, while both bind to catalytic Mg^{2+}; and third, both the DI of group II introns and the snRNA U5 of the spliceosome interact with the 5' exon under a similar structural arrangement (Figure 10) [183, 185, 197, 198, 199]. Furthermore, the protein Prp8 associates with spliceosomal RNAs and forms the catalytic core of the spliceosome, which is structurally similar to the IEP of group II introns (Figure 10) [183, 194, 198, 199, 200, 201, 202]. It was hypothesized that group II introns entered eukaryotic cells through endosymbiosis, and gradually evolved to the current form of non-LTR retrotransposons or spliceosomal RNAs by losing either the splicing or mobility activities respectively (Figure 10) [32, 93, 188, 203, 204, 205, 206, 207].

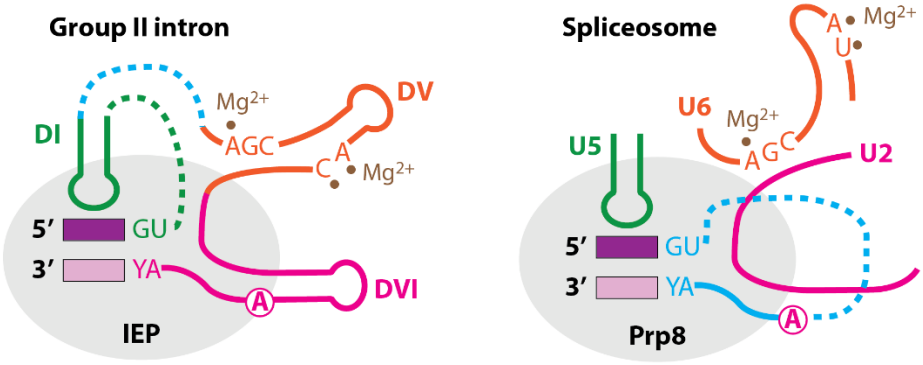


Figure 10. Comparison between group II introns and spliceosomes.

This figure highlights three major structural similarities between group II introns and spliceosomes: **1)** Both DI of group II introns and the snRNA U5 of spliceosomes have a stem-loop structure (green) that pairs with the 5' exon sequence (dark purple block). **2)** Both DV of group II introns and the snRNA U6 of spliceosomes have a catalytic triad (AGC) as well as an AC-bulge (orange) that bind to Mg^{2+}. **3)** The bulged A (pink, circled "A") is present in both DVI of group II introns (pink) or through pairing the snRNA U2 (pink) with the 3' end of the intron (blue) in spliceosomes. Figure adapted from references [197, 183].

Contributions of bioinformatics approaches

The term "bioinformatics" initially referred to the study of informatic processes in biotic systems, but more recently it has become a field combining various disciplines such as biology, mathematics and statistics, and computer science to analysis biological data [208, 209]. Contemporary bioinformatics is widely used as an alternative to traditional experiments and can generate data that may not be obtained through experiments. It has the ability analyse large datasets in the genomic or metagenomic scales in a relatively short time, and is capable of predicting genes of interest that exist in nature. As a subdivision of bioinformatics, phylogenetic tools can estimate the conservation and evolution among selected sequences. In addition, various models and algorithms have been designed to simulate molecular interactions in order to provide preliminary data to direct experimental design. In the rest of this section, several major divisions of the field of bioinformatics, especially those used in this dissertation, will be introduced along with examples of commonly used tools.

Sequence-related analyses

Sequence-related analyses focus on processing biomolecular sequences, such as DNA, RNA and polypeptides. They generally involve procedures of generating sequence alignments or searching for similar sequences, and the results can be used to compare sequences and identify mutations, identify specific genomic components, predict features and functions of related sequences. Alternatively, alignments can be used in phylogenetic analyses to trace evolutionary histories.

Two methods of generating sequence alignments are widely used. The first is pairwise sequence alignment, which compares two sequences at a time. BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) [210] is one of the most popular tool that uses pairwise alignment. BLAST is often applied to search for sequences based on identity from all sequences stored in a database. Identified matches are output along with a measure of statistical significance, which is usually an expectation value (E value) that describes the significance against random background noise. The lower the E value, the lower chance of having random sequences with a similar score in the database, and thus the more "significant" the match is. Therefore, the E value is often used to create a significance threshold for reporting results (BLAST FAQs,

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ).

According to the type of the input sequence (query) and the type of sequences stored in the target database, the standard BLAST contains four major programs: BLASTN (nucleotide against nucleotide), BLASTP (protein against protein), BLASTX (translated nucleotide against protein), and TBLASTN (protein against translated nucleotide). As a variant, PSI-BLAST (**P**osition-**S**pecific **I**terative-BLAST) is used to find proteins that are more distantly related to the query, which is a "profile" sequence and is generated by combining several closely related proteins. PSI-BLAST performs many iterations, and a new profile is generated by combining results after each iteration for the next round of search. In addition, with other modifications, BLAST can also be used to perform specialised searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

The second method is multiple sequence alignment, which compares many sequences at a time. This is more frequently used to align three or more sequences that are assumed to have evolved from the same common ancestor sequence. There are many commonly used programs that use different methods to create multiple sequence alignments, such as ClustalW [211], T-Coffee [212], MUSCLE [213] and HMMER [214]. Compared to the other three programs that use a progressive alignment method, HMMER uses the more complex HMM (**h**idden **M**arkov **m**odel) method and is capable of both creating a "profile" from aligned sequences and generating profile-based alignments.

Aligned sequences can be visualised for manual refinement. This function is supported by some sequence aligning programs as mentioned above, and there are also independent programs, such as SeaView [215] and Jalview [216], that can be used to provide graphic representation of aligned sequences. When a multiple sequence alignment is finalised, it can be used to generate a sequence logo that represents the sequence conservation and variation at specific locations, and the WebLogo server [217] is a common tool used to generate a sequence logo.

Structure-related analyses

Both nucleic acid and polypeptide sequences form secondary and tertiary structures to become biologically functional, and the prediction of these structures thus becomes an important topic of bioinformatics. For nucleic acids, one common approach is to predict

the secondary from the primary sequence, in which many possible structures are first generated, and the one with the lowest amount of free energy for folding is determined as the most stable structure. Available as both a web server and a stand-alone application, Mfold [218] is a popular tool for predicting secondary structures for DNA and RNA. Alternatively, the secondary structure can also be predicted comparatively, in which several homologous sequences are first aligned and then used to guide the prediction of structures for other homologous. One example that is capable of performing comparative prediction is program Infernal [219], which can create "covariance models" from RNA alignments. A covariance model resembles a "profile" generated by HMMER, but it is specialised for RNA, and thus can be used to identify conserved RNA secondary structures from primary sequence data.

Protein secondary structure elements, such as helices, sheets and coils, can be identified based on electrostatic forces and the association between certain residues and specific motifs [209], while the calculation of free energy is often involved to determine the most likely candidate. Likewise, comparative modelling based on similarity is another approach when the structure of a related protein is known. Many tools are available for analyses related to protein structures, such as the Phyre2 [220], I-TASSER [221] and ExPASy [222].

Phylogenetic analyses

As described earlier, aligned sequences are often used for phylogenetic analyses, which aim to infer the evolutionary history and relationships of individuals (taxa). During a phylogenetic inference, the input is most often aligned nucleic acid or amino acid sequences (molecular analysis), but can also be trait-based characteristics (morphological analysis). The result is a phylogeny, which is usually represented in a tree form (phylogenetic tree).

Many methods can be used to construct a phylogenetic tree. Distance-matrix based methods, in which pairwise distances are calculated from mismatches observed in the input alignment. Popular distance-based methods include **N**eighbour **J**oining (NJ) and **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic mean (UPGMA). They are easy to set up and the analysis can be done in a very short time. However, the trade-off is the results are not as accurate as other methods. Therefore, distance-based methods are often

used to build "draft" trees before any optimisation is applied. The second method is maximum parsimony (MP), which identifies the phylogeny that involves the smallest number of evolutionary events. The generated tree will be the most optimal under the assumption of minimal evolutionary changes; however, it is also sensitive to certain situations such as long-branch attraction, in which long-branched taxa are incorrectly placed together. Both distance-based and MP methods are relatively simple, and many sequence aligning programs have integrated functions to construct trees using these methods, such as PHYLIP [223], SeaView [215] and ClustalW [211].

The next type of methods is maximum likelihood (ML), which infers the probabilities for candidate trees. During the tree inference, a substitution model is required in order to estimate the probability (likelihood) of the sequences under the specified model in a particular tree. A substitution model describes sequence changes from one to another, and examples of commonly used models include GTR (**g**eneralised **t**ime **r**eversible) [224] for nucleotides and Dayhoff [225] for amino acids. The best-fitting model should be chosen for different datasets, and many programs also enable the user to implement their own models. Based on the maximum likelihood paradigm, Bayesian approaches provide another method that incorporates probabilities obtained previously, and generally uses Markov chain Monte Carlo sampling algorithms for tree inference. Both ML and Bayesian methods are considered more reliable compared to distance-based and MP methods, while the ability to specify the model is suitable to various conditions. However, both methods require intense computational resources and are more time-consuming. Example programs that are can construct trees using either or both methods are RAxML [226], PhyML [227], MrBayes [228] and MEGA [229].

A phylogenetic tree can be either rooted or unrooted. Rooted trees include a specific taxon is used to indicate the location of the most recent common ancestor of other taxa. In contrast, unrooted trees only infer the relationships among given taxa, but do not indicate the most recent common ancestor. A taxon that can be used as a root is usually an uncontroversial outgroup, which is known to be distinct from, but is still close enough to allow the meaningful comparison to other taxa.

The confidence level of estimated trees is evaluated after inference. Common methods are bootstrapping for distance-based, MP and ML trees and posterior probability for Bayesian trees. In practice, nodes with a bootstrap value of >= 70% (sometimes 75%) or a posterior probability of >= 95% is generally considered as valid, while nodes with lower

values are not considered as confident and may be left out from further analysis.

Data handling and database management

As being described so far, bioinformatic analyses generally involve processing the information of various sequences or molecules. Such biological information can be collected, organised and stored in databases to provide convenience for future access of specific data. One of the best-known databases is the GenBank built by NCBI (National Center for Biotechnology Information) [230], which contains many databases specific to different types of information (e.g. nucleotides, proteins). The GenBank is open to researches and enables the submission of experimental data by individual laboratories and projects. In addition, EMBL [231] and DDBJ [232] are other examples of databases for nucleotides. Similarly, UniProt [233], SWISS-PROT [234] and PDB [235] are databases specific for proteins, while PDB also contains information of solved tertiary structures of many biological molecules.

Databases are not only limited to primary sequences. For example, there are databases specific to functional motifs and families of protein or RNA molecules, such as CDD [236], Pfam [237] and Rfam [238]. Besides, some databases are specific for certain model organisms, such as EcoCyc [239] for *E. coli*, SGD [240] for yeasts, WormBase [241] for *C. elegans*, FlyBase [242] for Drosophila, ZFIN [243] for zebrafish, and Xenbase [244] for Xenopus.

In addition to those being introduced in this section, bioinformatics tools are widely applied in other topics such as analyses of gene expression, protein regulation, structural modelling and molecular interaction. However, since these topics are not involved in this dissertation, they are omitted in this introduction.

Aims of this dissertation

This dissertation includes four individual projects related to bioinformatic studies of bacterial retroelements. The first project described in Chapter II aimed to expand the current set of prokaryotic RTs and investigate their diversity, continuing from our 2008 publication [20]. As our lab focuses on group II introns, the next project described in Chapter III first aimed to update our collection of group II introns from GenBank. Based

on preliminary searches, it was expected that new classes could be revealed from the updated collection. Meanwhile, our current understanding of group II introns and their properties could be more systematically addressed, especially for each class. Many analyses were thus performed to all introns, and this project was done in collaboration with Ashley Jarding from our lab. Introns collected in this project contributed to the third project described in Chapter IV, which focused on the phylogeny of group II introns. This was a continuation of our 2009 study [181], which was limited by a small dataset. By taking advantage of the much larger dataset generated in Chapter III, this project aimed to solve questions left in the previous study, mainly about whether the IEP and RNA of group II introns coevolved, and how different classes are phylogenetically related to each other. The project described in Chapter V focused on a different retroelement, DGRs. Similar to the survey study for group II introns, this project intended to present a list of DGRs that are manually curated and can be used as references for future research.

Chapter II. Discovery of new reverse transcriptases in bacteria

Abstract

The first project of this dissertation aimed to perform an update to collect all putative prokaryotic RTs from the GenBank database as of July 2014. This was mainly done using an automated pipeline program previously published by our lab [245] with custom modifications. The majority of the new RT collection was occupied by group II introns, followed by retrons and DGRs. Almost all RT classes that were defined previously in 2008 [20] were expanded after the update. In addition, seven new RT classes, named unknown 10-16, were identified, although their functions could not yet be predicted. Considering the large expansion of the updated RT collection, it is believed that both the number and diversity of bacterial RTs will keep increasing in future searches.

Introduction

It has been known that prokaryotes contain RTs since the discovery of retrons [15, 153], and along with other types of RTs, such as group II introns and DGRs, it was indeed shown that there is a great diversity of bacterial RTs [20, 21]. Before the start of this project, a previous study carried out in our lab in 2008 identified 1049 bacterial RTs from GenBank using PSI-BLAST (**p**osition-**s**pecific **i**terated **b**asic **l**ocal **a**lignment **s**earch **t**ool), and these RTs were divided into 20 classes while 38 individuals were left unclassified [20]. While only group II introns showed clear evidence of mobility based on a high copy number in the same genome, other uncharacterised classes are thought to have other biological functions useful to the host, because they are of either a single or low copy number and exist in a limited number of host organisms [20]. Meanwhile, new types of RTs were expected to be identified with the ongoing, rapid expansion of sequence databanks. As an update to the study in 2008, the project described in this chapter searched the GenBank databases in July 2014 using an automated method. It first aimed to investigate by how much the new collection of bacterial RTs can be expanded, and also intended to identify any new classes of RTs.

Materials and Methods

Initial BLAST searches for putative full-length RTs

The search for new bacterial RTs began with a TBLASTN (protein query against translated nucleotide database) search. The query sequences included 311 RTs from the previous publications [14, 20], including RTs from group II introns, group II like 1-5 (G2L 1-5), AbiA, AbiK, DGRs, retrons, *rvt*, and nine unknown classes (Appendix A1). The search was against the GenBank non-redundant nucleotide collection (NT) as of July 2014, and the maximal E value (expect value) threshold was set to 1e-40 based on many preliminary test searches to eliminate less similar sequences. Custom Perl scripts (not provided in appendices) were used to perform BLAST searches in batch through the BLAST+ executables (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/) [246], and only hits from bacteria and archaea were retained.

Overlapping hits were merged into single entries to reduce duplicates and all hits were downloaded together with 10 kb flanking each side. The full-length proteins were predicted using the longest ORF matching the corresponding BLAST hit, and the ORF boundaries for some were adjusted to match available GenBank annotations. Based on the BLAST-generated alignment to the best hit, protein sequences that likely have premature stops and/or frame shifts were dropped from the dataset.

RT classification and class-wide alignments

Putative RTs were classified based on their closest relatives, which were determined through BLASTP (protein query against protein database). During the BLASTP comparison, each putative RT was used as the query and searched against a custom protein database consisting of ~1000 RTs from known groups plus 25 non-RT proteins as negative controls (Appendix A2) [20].

The top three hits for each query were used to classify each candidate RT, and if the search returned fewer than three hits, the RT was labelled as "N/A" indicating that not enough information was available for its classification. Three conditions were applied to classify candidates with at least three hits returned: 1) If all three hits or the first two hits belonged to the same type, the RT was assigned the same with high confidence; 2) If only two hits belonged to the same type and they were not of the top two, the RT was assigned the same but with lower confidence; 3) If all three hits were different, the RT

remained unclassified. Of note, at the end of the classification step, sequences labelled as "N/A" were all found to be truncated or have other problems (e.g. frameshifts) and were dropped from the dataset.

All except 87 RTs could be classified by the BLASTP-based method. As there might have been new classes among these 87 sequences, they were subjected to a pairwise BLASTP comparison together with 316 representative RTs from known classes. Every two RT sequences were compared using BLASTP, and based on the E value, they were either grouped to a known class, assigned as a new class or remained unclassified.

After the classification, sequences within each group of RT were automatically aligned by the program MUSCLE [213] except for group II introns, which were aligned by the program HMMER [214] using a sequence profile previously established by our lab. These automatic sequence alignments were refined manually, and truncated sequences were identified and excluded from the dataset.

In addition, it was later noticed that four sequences identified from the previous study [20], including three from G2L5 and one from unclassified G2L, were missed in this update because GenBank had removed their entries (reasons were not stated). To maintain the completeness, these four sequences were appended into the new dataset. The complete flowchart of this updated RT collection is depicted in Figure 11.

Domain composition analysis

The domain composition for each RT was analysed by Pfam [237] using default settings. To reduce the processing time and the server load, sequences that were highly similar to our group II collection were not subjected to this analysis. The search was performed in batch through a custom Perl script (not provided), and results were verified by submitting to CDD (Conserved Domain Database, https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) [236].

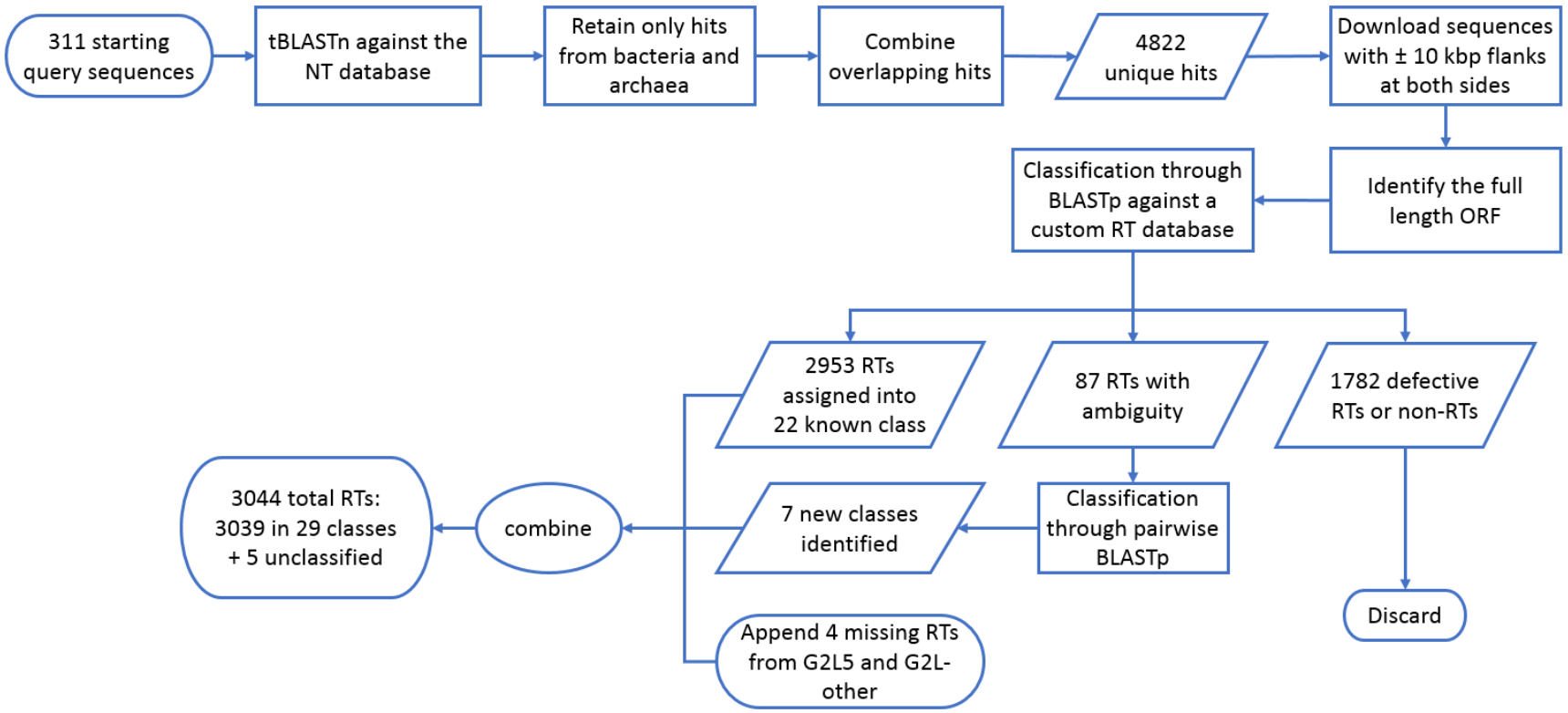


Figure 11. Workflow of collecting bacterial RTs from GenBank.

The search began with a TBLASTN search against GenBank's NT database using 331 RTs as the query. Only hits from bacteria and archaea were retained, and overlapping hits were merged to unique entries followed by sequence downloading including ± 10 kb flanks. Full-length ORFs were identified from the downloaded sequences and classified first based on closest relatives through BLASTP. Defective RTs and non-RTs were excluded. Remaining RTs were classified into 22 known classes leaving 87 RTs unclassified, which were further classified by pairwise BLASTP into seven new classes. At last, with the addition of four missed sequences from the previous collection, the dataset contained 3044 RTs in total.

Results and Discussion

The expanded collection of bacterial RTs

The main purpose of this study was to collect bacterial and archaeal RTs from the most recent GenBank database as an update to the last study performed in 2008 [20]. Previously, RTs were detected by PSI-BLAST, which was manually performed for multiple iterations. During this update, a different approach was used by searching for RTs from GenBank through TBLASTN, which is easier to automate (Figure 11). Excluding defective sequences, this update detected 3040 putative bacterial RTs that were later classified into 29 classes, while seven of the 29 classes were newly revealed from this study. All sequences were verified to have a complete RT domain through both Pfam and CDD, whereas an RT domain was either missing or incomplete in the set of excluded sequences. Therefore, the automated approach used for this update is capable of detecting functional RTs and requires less manual effort compared to the PSI-BLAST-based method used in the previous study.

This updated dataset showed a greater diversity of bacterial RTs by revealing seven new classes. But the overall distributions of RTs from different classes is similar to the previous result: the majority of RTs are still from group II introns (75%), followed by retrons (12%) and DGRs (3%) (Figure 12). The remaining 10% fell into 26 classes, including Abi-associated RTs (AbiA, AbiK, Abi-P2), group II like RTs G2L1-4, the *rvt* elements, unknown groups 1-16 and five unclassified individual sequences (Figure 12). The class of *rvt* elements, which was separately reported in 2014 [14], was new to the previous collection but was not a new class revealed in the update; while unknowns 10-16 were new classes revealed at this time.

Five individual RT sequences still remained unclassified. They were compared to GenBank's non-redundant protein database (NR) and showed less significant E values ranging between 1e-26 to 1e-6, indicating they are unique to both the other RT sequences in this collection and also the NR database. However, it is believed many of them will be classified in the future when more sequences are added into GenBank. For example, the group named unknown 10, which contains five members, used to be an unclassified individual in the previous collection [20].

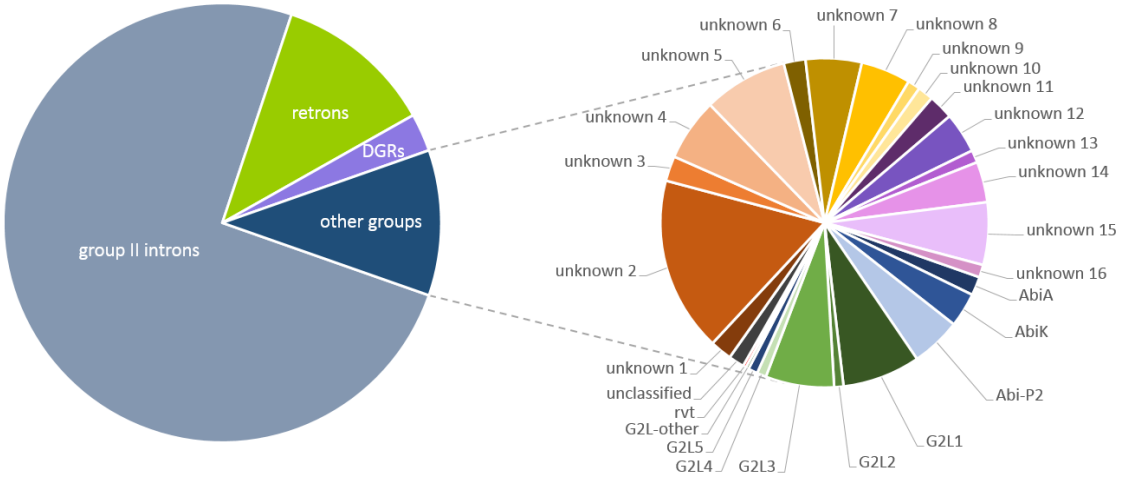


Figure 12. Class distribution of RTs from this collection.
The majority of the set is from group II introns, retrons and DGRs, while the remaining ~10% showed a higher diversity that fell into 26 classes plus five unclassified individuals.

Of the new RT collection, the group sizes of most classes increased except for G2L2 and unknown 1, which both are smaller compared to the previous collection by one sequence. Besides, another two classes, G2L5 and G2L-other (unclassified G2L) were found completely missing in the new collection. In the previous collection, G2L5 and G2L-other were both small groups and contained three and one sequences, respectively. To verify whether they were missed due to a different method being used, additional BLASTP searches were performed using all four sequences (ZP_01851752, ZP_01854760, ZP_01090701, ZP_01872295) as queries. However, no hit was returned by the search, while the four entries themselves were noticed to have been removed from the GenBank database, although the specific reason was not stated. In order to maintain the completeness of the collection, these four missed sequences were appended to the new RT collection (Figure 11). The removal of entries from GenBank (e.g. when a more recent record is published, or when the whole genome where the old entry belonged to was completely sequenced) could have been the reason for G2L2 and unknown 1 to be smaller. However, this was not further investigated because the new and previous collections only differ by one sequence, which is not significant.

Domain composition and functional prediction of RTs

To illustrate the composition of the RT domain for each class, the RTs were first aligned to the IEPs of group II introns, which contain all RT motifs 0-7. It was noticed that not all RTs could be aligned along the entire RT domain, even though that almost all classes could be aligned across RT motifs 3-6 (Figure 13), which are essential to form the active core and correspond to the finger and palm subdomains as observed in other polymerases [247, 248]. This indicates that all RTs identified in the updated collection share a similar core structure.

Next, whether other RT classes would also contain the thumb domain was investigated. In group II intron IEPs, the X domain is believed to be equivalent to the thumb domain, and is downstream of the RT domain. However, the thumb domain could not be predicted for most other RT classes except for a few G2L classes due to the lack of sequence similarity [249]. Regardless, almost all of the remaining classes except for unknown 11 contain an extension of at least 50 aa downstream of the RT, which is long enough for a thumb domain [250]. Therefore, together with the presence of the finger and palm domains, most RT classes are likely functional polymerases even though the thumb domain could not be predicted from sequence similarity.

In addition to the RT and thumb domains, the function of an RT-containing protein could be inferred if other domains are present. Among all classes, only group II introns clearly contain an EN domain, which facilitates their retrohoming process. Some classes of RTs have already been characterised previously, such as that G2L1 and G2L2 that are associated with some CRISPR/Cas systems, and AbiA, AbiK and Abi-P2 that are related to the phage immunity Abi system [20]. However, functions of most other RT classes remained unclear. Therefore, additional protein motifs were analysed using Pfam and CDD for all RT classes in order to make predictions for each class, and only domains present in over half the members of a class were considered as a class-specific feature (Figure 13).

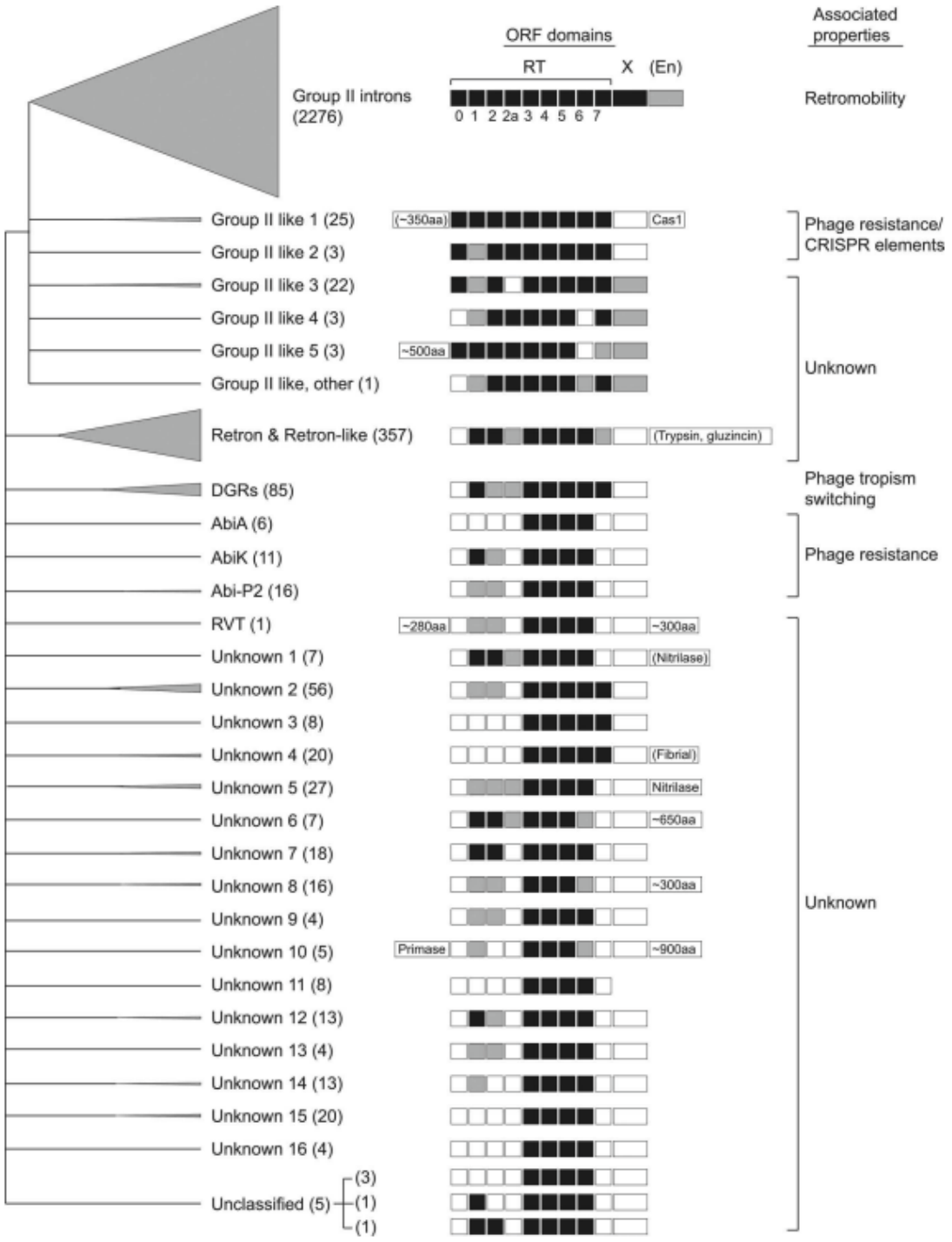


Figure 13. Domain composition of each RT class.

Classes are depicted in a tree format that places similar classes (group II and group II like) into the same clade. The size of triangle before each class name reflects the class size, and the number of each class is given in parentheses after the class name. Sequences are depicted by squares representing individual RT motifs and other domains or extensions. Black, grey or white indicate clear, ambiguous or absence of alignability to group II introns respectively. Predicted properties of each class is listed on the right.

Only a limited number of protein motifs could be detected. Two subsets of retrons (based on a preliminary neighbour-joining tree, not provided here) showed a trypsin or gluzincin domain, which are both associated with proteinase activity, at the C terminus, while the rest of retrons had no additional detectable domains. This indicates some retrons may be related to protein digestion, and also implies that retrons may have different biological functions depending on the host, supporting a previous hypothesis [154]. Unknown groups 1 and 5 have a nitrilase domain, which has been characterised to break C-N bonds in nonpeptide cleavages involved in small molecule metabolism [21, 236]; and unknown group 4 has the pilus-related domain fimbrial [20]. Regardless, these motifs are not directly involved in normal RT functions, and it is impossible to predict whether they contribute to the RT without experiments. Unknown 10 was detected to contain a primase domain (Figure 13). Primases are a class of RNA polymerases that are essential during DNA replication since they synthesize short RNA primers to initiate the process of DNA synthesis [251, 252, 253, 254]. Therefore, it would be reasonable to hypothesize that the primase domain involved in an RT could assist the process of reverse transcription.

Aside from additional protein motifs, a different approach to predict functions of RTs was based on genes in the neighbourhood of the RT [21]. Therefore, information about flanking genes for each RT sequence in this collection was gathered from available GenBank annotations. This did not yield informative results (data not provided), as most genes in the flanks were annotated as hypothetical proteins or transposon-related genes. Although a few RTs are located close to some housekeeping genes, none of these housekeeping genes was consistently present in the neighbourhood of all RTs from the same class, and thus no clear pattern could be concluded to predict the function of the RT.

Conclusion

Bacterial RTs are clearly diverse and widespread. As anticipated, this project increased examples of known types of RTs in bacteria and revealed seven new classes (Unknowns 10-16). Moreover, based on the domain composition, all RT classes are expected to be functional, although a few may not function as classic polymerases (e.g. AbiK). However, the specific function of each class of RTs remains unclear for most

classes due to the lack of information about additional functional motifs or the properties of genes in the neighbourhood of the RT.

In the technical aspect, the automated approach used in this project is believed to be capable of detecting putatively functional RT sequences. It mainly differed from the previous study by using regular BLAST searches instead of PSI-BLAST [20], which was designed to identify proteins that are more distantly related. In spite of using a different approach, all RT classes discovered previously have been detected in this project, indicating that regular BLAST, which is easier to be automate than PSI-BLAST, is sufficient. On the other hand, because some proteins with the same function vary in sequences, tertiary structure-based protein identification could be considered in the future to reveal novel RTs that are less similar in sequence, although this would be limited by available protein structures.

Chapter III. Learning about group II introns through bioinformatic analyses

Abstract

Group II introns are retroelements found in bacterial and archaeal genomes as well as in many organellar genomes. The aim of the project described in this chapter was to create a comprehensive collection of group II introns in bacteria and archaea from data available in the GenBank database. Full-length intron sequences were identified using an automated method, and their secondary structures were first folded automatically followed by manual correction, which was mainly based on the consensus structure of each class. This updated intron collection is about 4-fold larger than the old dataset stored in our group II intron database (http://webapps2.ucalgary.ca/~groupii/). All compiled introns were then subjected to a series of analyses to update our current understand of group II introns, including the following topics: 1) A phylogenetic tree revealed a new IEP-based Class G and a tentative new Class H; secondary structures were then used to create the consensus RNA structures of these two classes, while the consensus structure for Class A was also refined since this class was greatly enlarged in the update. 2) Target specificity was analysed manually for a small subset of introns, and a preference of homing site could be observed from all classes. 3) Biases were detected for introns inserting into the leading or lagging strand, which was generally consistent to the hypothesis that EN-lacking introns more often rely on a priming site provided by the lagging strand. 4) A variety of non-standard intron organisations were identified, including twintrons, tandem introns and more complicated situations. Multiple copies of the same "intron complex" could be observed for many examples, indicating these intron complexes may still be capable of splicing and not merely inactivated remnants. All newly identified introns were stored in our group II intron database, which also underwent technical improvements for a better performance.

Introduction

As a type of mobile element, group II introns are widespread in bacteria and sometimes found in high copy numbers. Our lab maintains a database for group II introns (http://webapps2.ucalgary.ca/~groupii/) that contains a large number of putatively functional group II introns from bacteria and archaea. This database was initially established in 2002 and contained about 40 introns [255]. In 2012, a pipeline program was developed by our lab to identify group II introns automatically from large DNA datasets, such as the GenBank database [245]. Since then, the number of new introns has continuously increased and they have been added to the database [256]. During previous searches using the pipeline program, it was noticed that the same intron sequence may exist in different genomes. To avoid recording such duplicates, our database only keeps one record among multiple introns that are highly similar (>= 95% identical in sequence) to each other and belong to the same host species. This intron representative was thus referred as the intron "prototype". By the end of 2013, the number of intron prototypes stored in our database was about 700.

However, a few issues were also found from the pipeline program. First, introns from Class E are often mislabelled as unclassified due to inconsistent file names in the source code, and these introns were often discarded by the pipeline in subsequent steps. Second, even though the total number of introns has increased greatly since we began to identify introns automatically, corresponding changes have never been maintained up to date, including sequence profiles used for classification or secondary structures used for automatic intron folding; and this may cause incorrect predictions due to the use of old profiles. Third, the pipeline program interacts directly with other programs, such as the BLAST+ suite [246], HMMER [214] and Infernal [219], and updates of those external programs require the pipeline to be updated.

In addition to the issues mentioned above, there are many limitations of the pipeline [245]. First, the prediction of RNA boundaries may be ambiguous for some introns from CL1 and Class B due to the presence of a 5'- or 3' extension [257, 258]. Second, since the pipeline begins with searching for group II-related RTs, introns without an ORF (ORF-less) would not be detected, even though they are rare compared to standard introns [259, 260]. Third, this pipeline was designed to discover standard introns only, and thus twintrons (intron within another intron), tandem introns and other non-standard introns would not be revealed.

The first aim of the project described in this chapter was to solve currently identified issues of the pipeline by either modifying the source code, updating associated files or providing alternative approaches. Next, because the last search for a comprehensive set of introns was done in 2013, which was four years before I started to update the pipeline program, the fixed pipeline program was thus expected to reveal an even larger set of group II introns. Therefore, the second aim of this project was to perform several bioinformatic analyses in order to update our knowledges in group II introns based on the most recent dataset, such as from an updated phylogenetic tree, revisions for RNA consensus structures, specificities of homing sites or insertion strands, and detection of non-standard but possibly active intron complexes.

Materials and Methods

As a clarification, this project was performed in collaboration with Ashley Jarding in our lab. Typically, I performed all computational work such as writing scripts for batch data processing, but the rest of works, such as intron folding and other data analysing, were done with Ashley Jarding, and her contribution will be specified.

Intron collection from GenBank

Two updates were performed in 2013 and 2017 to search for introns from GenBank, and both used the same pipeline framework [245]. All scripts used in the 2013 update were written by Abebe *et al.* [245], and some scripts used in the 2017 update were either modified or revised by me in order to solve issues described in the introduction (source codes not provided).

Detection and classification of putative group II RTs

The search began with identifying group II-related RTs through BLASTX (protein query against nucleotide database) searches against GenBank's non-redundant nucleotide database (NR). A total of 42 intron representatives selected from each class (bacterial A-G, CL1, CL2, ML) plus six unclassified introns were used as query sequences. Overlapping hits were merged to form a list of non-overlapping "unique" hits, and the

longest ORF contained each hit was used as the putative IEP. Nucleotide sequences of each putative IEP, including ± 3000 bp flanking each side, were downloaded for later analyses.

False hits (non-group II RTs) were identified through BLASTP by comparing each hit to a set of known RTs including both group II and non-group II RTs (e.g. DGR, retron), and if any of the top three hits belonged to a non-group II RT, the hit was dropped from the dataset. Remaining sequences were classified through another BLAST comparison against representatives selected from each class. If at least two of the top three hits belonged to the same class, the hit was also assigned to the same class. Otherwise, the putative IEP remained unclassified.

Filtering for incomplete group II RTs

To only retain RTs that are likely to be functional, candidate IEPs were evaluated for their domain completeness. This was done through a BLASTP-based method. First, a reference RT was selected from existing functional IEPs for each class. For each of RT motifs 0-7, domain X and domain EN (if present), an amino acid was selected from the most conserved area as an "anchor" (Figure 14 top, Appendix B1). These anchors were used to verify whether the corresponding domain or motif is present in a candidate. Each candidate RT was compared to the reference RT from the same class through BLASTP. If all anchors are aligned to an amino acid of the candidate, the candidate IEP is thus considered to have all domains and thus putatively functional. As illustrated in Figure 14, the reference RT is shown on top with two example anchors marked in pink and blue respectively. Two sample sequences 1 and 2 are aligned to this reference. Sample 1 is considered as complete because it contains an amino acid alignable to both anchors, while sample 2 is considered incomplete because it does not have an amino acid alignable to the anchor marked in blue. Incomplete RTs (one or more anchors are not aligned to an amino acid from the candidate) were excluded from the dataset.

Figure 14. Example of using amino acid "anchors" to evaluate the completeness of candidate RTs.

The representative RT is selected from each class, and for each domain and motif (RT motifs 0-7, X and EN), one amino acid was selected as the "anchor" (pink and blue). Candidate RTs were compared to the representative through BLASTP. If the candidate shows an amino acid alignable to the representative at all anchored positions, it is considered complete (Sample 1). If the candidate lacks an amino acid alignable to at least one anchored position, it is considered incomplete (Sample 2) and excluded from the dataset.

Prediction of RNA boundaries and folding of intron secondary structures

As group II intron RNAs generally lack sequence similarity but can still be structurally aligned, the program HMMER was used to generate sequence profiles from currently available RNA alignments for each class, and the profiles were used to predict the intron RNA boundary. Profiles for each RNA domain were separately created. The entire domain IV was excluded because it mainly contains the ORF and is not essentially required to form a functional RNA tertiary structure. Sequences downloaded in the previous step (RT with ± 3000 bp flanks) were searched for RNA segments matching these profiles using HMMER. Sequences that showed matches to all five domain profiles were forwarded to an automated folding script written by Michael Abebe (not provided); this script first align new RNA sequences to a user-defined RNA alignment through Infernal [219], then generates folding constraints and folding new sequences through Mfold [218]. Sequences that lacked one or more matches of these five domains were dropped from following analyses. Finally, all folded RNA structures were aligned to the current RNA alignments and with manual refinements. Introns that could not be folded into a standard six-domain structure were excluded. The step of intron folding and correction was done together with Ashley Jarding.

Detection of non-standard intron complexes

The pipeline program was not designed to find non-standard introns due to their complexity in organisation. Therefore, such introns were identified through visual inspection, and this was done together with Ashley Jarding. All previously downloaded

sequences were subjected to finding such non-standard introns. Boundaries of potential IEP and RNA, which were predicted in previous steps by either BLAST or HMMER searches, were visualised using a custom Perl script (not provided) as demonstrated in Figure 15.

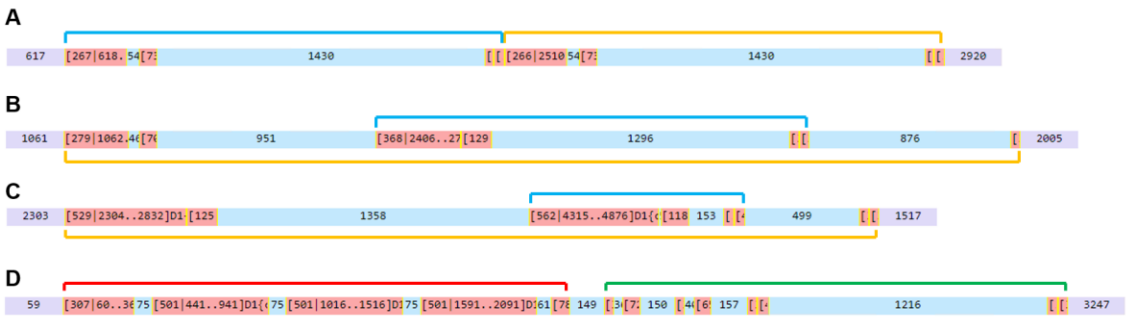


Figure 15. Examples of visual inspection for non-standard introns.
Blue and red blocks (with yellow borders) are drawn to scale and correspond to IEP and RNA matches performed by HMMER respectively. At both sides, purple blocks indicate flanking areas with no matches, and are also not drawn to scale. Numbers written on each block indicate the length in nucleotides. Red blocks also contain more details including the boundary coordinates and the target domain it matches, but is truncated in this display but can be expanded when the image was opened in the original format (HTML). The visual inspection focused on the arrangement between different blocks. **A)** Two introns in tandem, while the blue and orange bars correspond to the upstream and downstream copies respectively; **B)** A classical twintron, while the blue and orange bars correspond to the inner and outer copy respectively; **C)** An ORF-less intron indicated by the blue bar, and meanwhile, this is probably the inner intron of a twintron, in which the outer copy is indicated by the orange bar; **D)** A non-standard example of four introns nested into each other, while the red bar corresponds to repeated domains of DI-III, the green bar corresponds to repeated domains of DV-VI.

The visual inspection mainly relied on the order of RNA segments, and mainly focused on finding three types of intron complexes: tandem introns, twintrons and ORF-less introns. If using "D5" to represent RNA domains DI-III (domains at the 5' end) and "D3" to represent RNA domains DV-VI (domains at the 3' end), a standard intron can be written as "D5-D3", in which the IEP resides between D5 and D3. Therefore, two introns in tandem can be written as "(D5-D3)-(D5-D3)", in which a complete intron is indicated in parentheses (Figure 15A). Similarly, a typical twintron can be written as "[D5]-(D5-D3)-[D3]", in which the inner intron is indicated in parentheses and the outer intron, which is interrupted by the inner intron, is indicated in square brackets (Figure 15B). ORF-less introns should generally have the same organisation as standard introns ("D5-D3"), but

lack the IEP portion (Figure 15C).

In addition, other more complicated organisations were identified during the visual inspection. Such intron "complexes" often contain multiple layers of introns but segments can form standard introns if each spliced in order (Figure 15D). All non-standard introns were also folded using the same procedure as described in the previous section.

Phylogenetic inference based on the IEP sequences

The program RAxML [226] was used to construct phylogenetic trees using all IEP sequences under the LG model and with 1000 bootstrap replicates. The IEP sequence alignment was produced by HMMER with manual refinements, and the global lenient IEP mask (as described in Chapter IV) was used to include only the more conserved sites.

Automatic prediction of intron insertion sites

BLASTX was used to predict automatically the intron insertion sites. The query sequence was the ligated exon sequences (~300 bp at each side), and the target database was GenBank's NR database. Of all proteins matched in the result hits, only matches across the ligated junction without a frameshift were assigned as the predicted host gene of an intron. Custom Perl scripts (not provided) were used to perform searches in batch and screen the results.

Analysis of the strand preference of intron insertion

Only introns belonging to a complete genome (based on the title of GenBank records) were included in this analysis. The origin and terminus of DNA replication of each genome were predicted based on the cumulative GC-skew [261, 262, 263] using the lowest and highest values respectively. The GC-skew was calculated by a Perl script (available at https://github.com/Geo-omics/scripts/blob/master/AssemblyTools/gcSkew.pl) with the window size set to 1000 bp. Because the origin and terminus are on average away by 50% of the entire genome size, manual spot checks (done together with Ashley Jarding) were performed using application GenSkew (http://genskew.csb.univie.ac.at) to find genomes that do not have

a clear GC-skew (Figure 16). It was decided to include only genomes that had their termini predicted to be 50 ± 10% of the whole genome size away from the origin. Meanwhile, to avoid highly similar genomes, pairwise BLASTN was used to pick up potential duplicates that have a >=98% sequence identity over the full-length of the intron and ± 2 kb of the flanking regions.

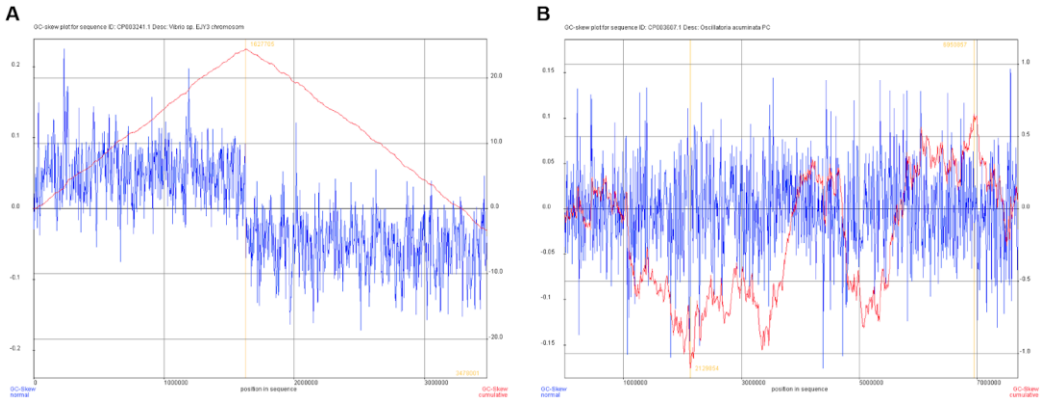


Figure 16. Example genomes with clear and ambiguous GC-skews.
The blue line corresponds to the normal GC-skew within specified window size of 1000 bp, and the red line corresponds to the cumulative GC-skew. **A)** Genome CP003241.1 shows a clear GC-skew, in which positions with the lowest and highest cumulative CG-skews were predicted to be the origin and terminus respectively. **B)** Genome CP003607.1 shows an ambiguous GC-skew and was excluded. Images were generated by application GenSkew.

Results and Discussion

The expanded collection of group II introns and new classes

Before the two updates, there were 329 intron prototypes stored in our database, and they fall into nine classes: Classes A-F, CL1, CL2 and ML. The two updates in 2013 and 2017 revealed 1179 new intron prototypes in total, which is about 4-fold larger than the old set. Of all the newly predicted introns, 946 are putatively functional, and the remainder 221 have a defective IEP but an apparently functional RNA. Together with the 329 introns in the old dataset, there are 1275 putatively functional intron prototypes in the final dataset. Compared to the old dataset, all known classes have been greatly increased, especially Class A, which now contains 51 members compared to the initial number of three (Figure 17).

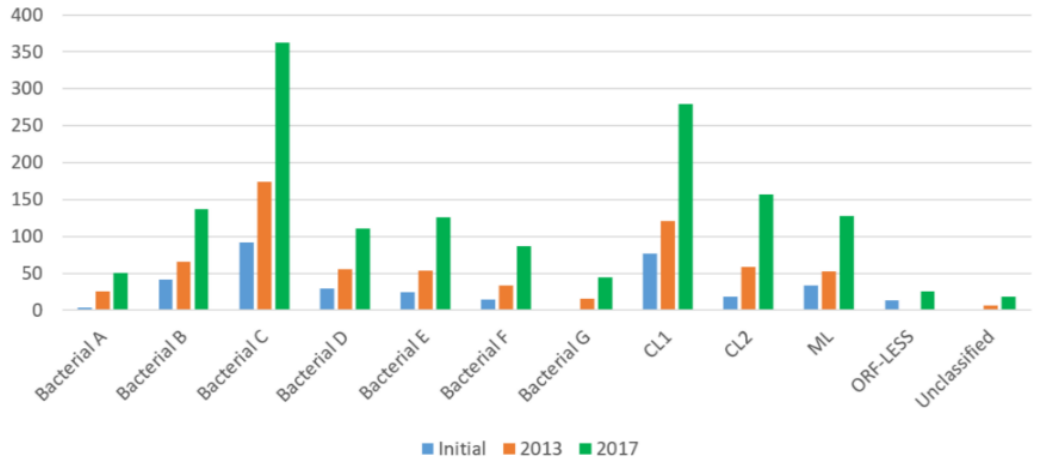


Figure 17. Increase in total number of introns of each class.
Comparison is between the initial collection before 2013 and after the two updates performed in 2013 and 2017. Note that Class G is a newly discovered class and did not exist in the initial dataset before 2013.

In order to clearly classify all newly identified introns, a phylogenetic tree was constructed using the IEPs of all 1275 intron prototypes (Figure 18). Most classes are supported by forming a single clade with a >= 75% bootstrap value except for Classes C and F, which formed into single clades but without supports (Figure 18). A new class, named Class G (Figure 18, red), was revealed by the IEP-based phylogenetic analysis, and it contains 45 introns in total (Figure 17).

A few introns remained unclassified based on the tree, and most of them scatter between Classes C and F (Figure 18, black), except for one intron, Hp.au.I3, that is close to the cluster of A+B+ML (Figure 18). Of these unclassified introns, a small clade of four individuals was observed with support, which then led to the tentative establishment of another new class, named Class H (Figure 18, light green).

After confirming the classification for all introns, intron sequences of the same class were structurally aligned according to their folded RNA structure, which was then used to update the RNA consensus structure for each class. However, it was noticed that even with a larger number of introns, the currently confirmed RNA consensus structure was not affected for most classes. Therefore, updates were only made for Class A (Appendix C), which had increased from three to 51 members. RNA consensus structures were also created for the two new classes, G and H (Appendix C).

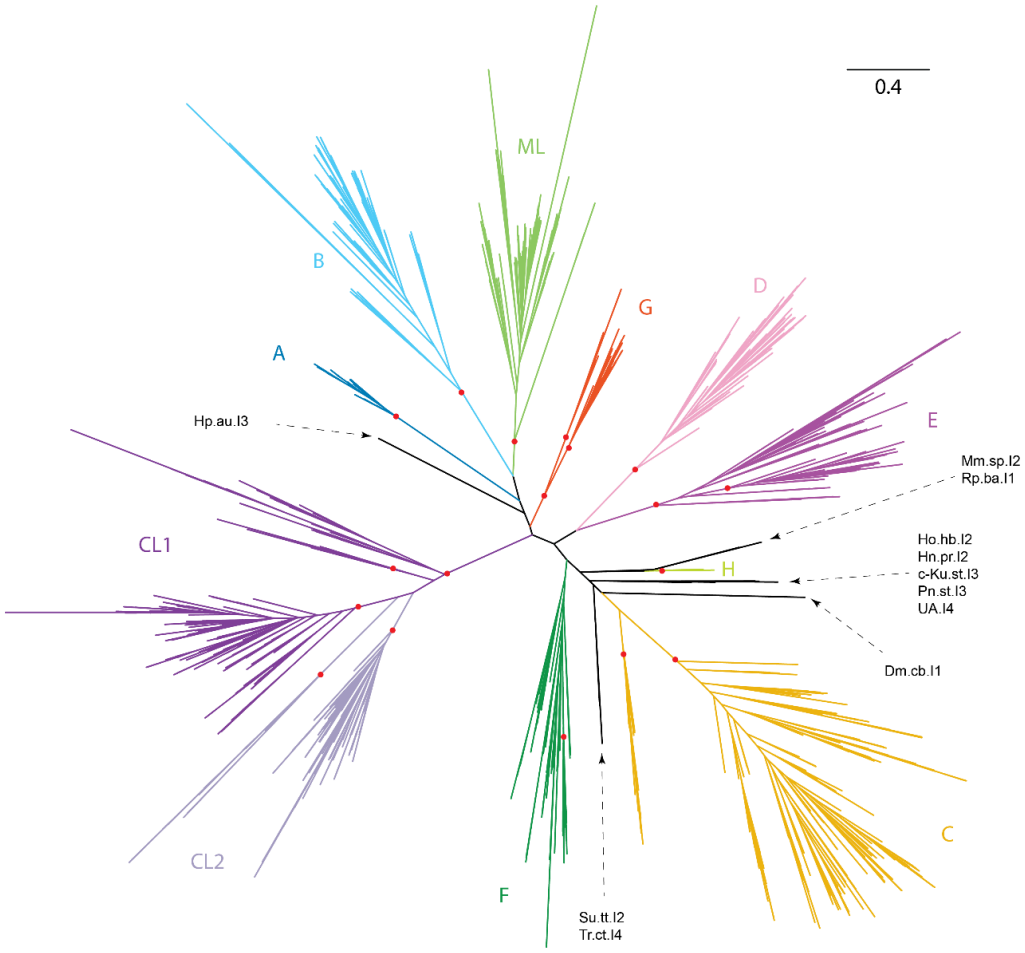


Figure 18. The IEP-based phylogenetic tree of 1275 introns from the most recent compilation.

Classes are colour coded and strongly supported (bootstrap >= 75) major clades are indicated by red dots. Names of all unclassified introns are pointed out by dashed lines. The tree was made using the program RAxML with 1000 bootstrap replicates.

Intron insertion patterns

Because the phylogenetic tree clearly separates all classes from each other except Class F, which still form a single clade, members of each class should be more phylogenetically related to each other compared to members from other classes. Therefore, it would also be more likely that members of the same class tend to target more similar homing sites. To investigate this, the homing sites of a subset of introns (~200) were analysed manually by examining whether the ligated exon sequences belonging to a known gene or potential ORF through BLASTX. Not all introns were subjected in this analysis due to large amount of manual work. This work was contributed by all members in the Zimmerly lab in 2013.

As expected, there is a trend that each class has a preference in its homing site (Figure 19A): Classes A, D and G mainly target transposon-related genes, especially Class A that has all members inserting exclusively into transposases; Classes B, ML, CL1 and CL2 have a sizable number (25-50%) inserting into various housekeeping genes; and Classes C, E, F as well as most unclassified introns were mainly found intergenic, while Class C was the most unique as it mainly targets the short region downstream of a transcriptional terminator. In addition, many introns were found inserting into hypothetical proteins, which may be unidentified housekeeping genes but without enough evidence to be characterised conclusively as such. However, even when such hypothetical proteins were considered as housekeeping genes, the overall pattern of homing sites for each class was not affected.

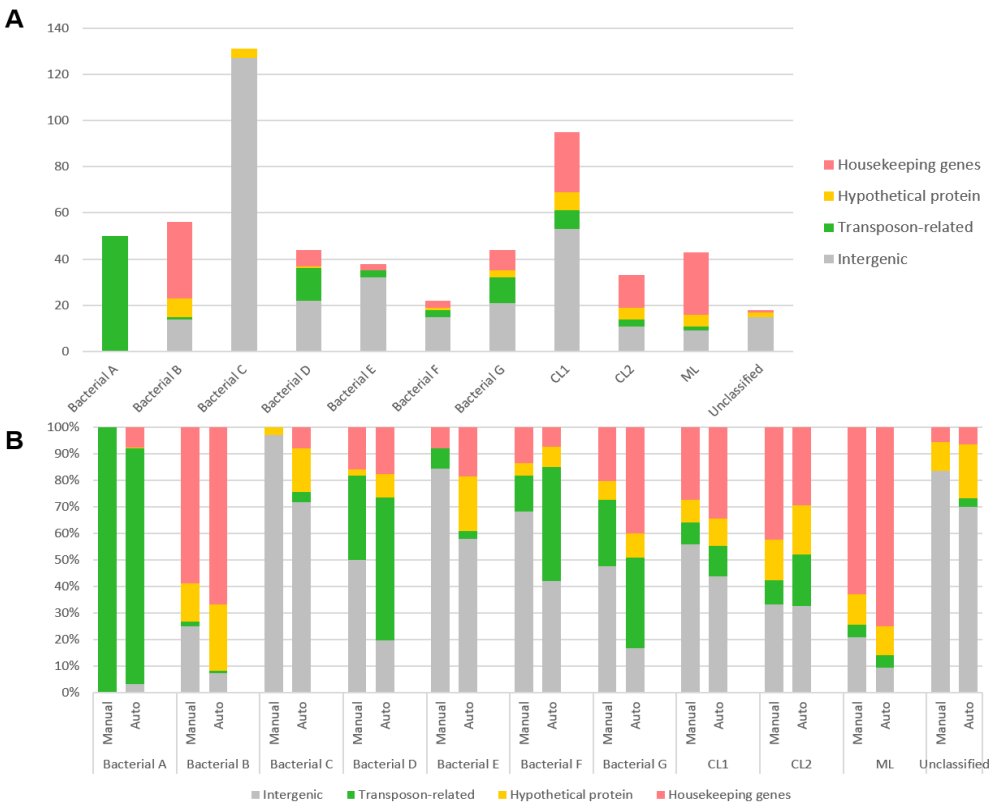


Figure 19. Distribution of insertion sites for each class.

A) Insertion patterns through manual inspection using a small subset of introns. The y-axis corresponds to the absolute number of introns. **B)** Comparison of the type of insertion site between the manual and automated approaches, while the y-axis corresponds to the proportion in percentage of each type. Data of the manual inspection is the same as shown in A), and data for the automated inspection was based on automated BLASTX searches of about 6000 intron copies.

The manually examined subset was believed to represent the full collection of group II introns. However, during the process of identifying multiple copies for each intron prototype, it was also noticed that some introns could have different homing sites for different copies. Therefore, an automated approach was used to predict the insertion sites for all identified intron copies (~6000 introns). This was done by comparing the ligated exons against GenBank's protein database using BLASTX, and top hits that covered the junction of exons with no frameshifts were retained as the predicted homing site.

Figure 19B compares results between the manual examination and the automated approach for each class. Overall, both methods showed the same trend in preference for homing sites, but there were also some variations that could not be considered as minor. For example, the entire Class A was confirmed to insert into transposases through the manual examination, but the automated method predicted some inserting into housekeeping genes (Figure 19B); the manual examination showed almost the entire Class C was inserting into intergenic areas except for a very small portion (~3%), but the automatic prediction showed a rather large portion (~30%) of Class C introns could be inserting into genes (Figure 19B).

To verify whether such disagreements reflect real differences caused after increasing the dataset, affected classes were spot checked. As a result, all sampled individuals were found to be incorrectly predicted by the automated approach. The first reason was inaccurate GenBank annotations; and the second reason was incorrect BLAST hits, which usually occurred when the intron inserted at the very beginning or end of its host gene. In the latter case, the real host gene may not always be the best hit, causing the prediction of its host gene to be incorrectly assigned. In contrast, manual examination spends more time to verify all top hits for each ligated exon sequence, and thus ambiguous GenBank annotations can be replaced by more reliable information, and the "real" best hit can be justified by comparing all other hits. Therefore, although the automated approach may never be as accurate as manual inspections, it is much faster than manual inspection, has the ability to process large datasets, and can still lead to the same trend. This automated approach is expected to be improved by considering more hits returned by BLAST to achieve more accurate predictions, which will gradually replace the labour-demanding manual inspection.

The relationship between intron host organisms and IEP-based classification

Aside from the direct analysis of identifying the ligated exon sequences, a different analysis showed that the host organisms of introns appear to correlate well with the IEP-based tree (Appendix B2). Sometimes, almost an entire class is found in a single phylum. For example, all Class A introns except one are from proteobacteria, all Class B introns except one are from firmicutes, and almost all CL2 introns are from a cyanobacterium host (Appendix B2). In contrast, some classes are mixed with several phyla. For example, Class G introns split into two clades that are mainly from proteobacteria and actinobacteria respectively, while ML introns are mainly from firmicutes, proteobacteria and Bacteroidetes (Appendix B). Very often, it was also observed that introns belonging to the same phylum cluster into clades (Appendix B).

Altogether, this correlation explains that introns from the same class are more likely to target the same or similar homing sites. On the other hand, this also implies a possible direction of horizontal transfer. As hypothesized in previous publication [264] as well as in Chapter IV of this dissertation, Class C is the most likely to be the ancestral class among all group II introns, while the majority of earlier branching introns of Class C are from firmicutes, while the rest are mainly from proteobacteria. According to the 16S rRNA-based phylogenetic tree published by 'The All-Species Living Tree' Project [265, 266], firmicutes branched earlier than other phyla that contain group II introns (e.g. actinobacteria and proteobacteria). However, this project did not examine this topic in detail. Future analysis can thus further investigate the connection between the IEP-based phylogeny and the evolution of involved host organisms, which could hopefully explicate the distribution of introns among diverse species.

Bias in the strand of intron insertion

As mentioned earlier, the EN domain of the IEP is responsible for cleaving the bottom strand of double stranded DNA and enables reverse transcription of the intron RNA during retrohoming. Some introns lacking the EN domain have still been shown to be capable of retrohoming by using an alternative primer coming from the nascent lagging strand during DNA replication [132, 134, 267]. As a result, such introns would preferably insert into the leading strand.

To verify this hypothesis, the insertion strand of all introns belonging to a completely

sequenced genome was analysed. First, the origin and terminus of each genome were estimated based on the GC-skew [261, 262, 263] and the portion of leading and lagging strands were thus predicted based on the location of origin and terminus. Genomes that did not show a clear GC-skew were excluded as described in Methods (Figure 16). Because it is common for different substrains of some organisms (e.g. *E. coli*) to be repeatedly sequenced, these genomes may be nearly identical in sequence and contain identical introns inserted at the same site. Such intron duplicates may cause an artificial bias if included in the analysis as they do not represent independent mobility events. Therefore, in order to identify and eliminate them, all introns including 2 kb of flanks at both sides were subject in pairwise BLASTN-based comparisons, and sequences that are >= 98% identical were considered duplicates and excluded. The final dataset included 2530 unique intron copies from 911 completely sequenced genomes. Within each class, the number of introns inserted at each 1% interval of the whole genome was plotted as shown in Figure 20.

As an average for most bacterial genomes, while setting the origin at 0 and 100% positions of a genome, the 0-50% region on the top strand and the 50-100% region on the bottom strand will correspond to the theoretical leading strands during DNA replication, and the rest of the regions will be the theoretical lagging strands. Therefore, an obvious bias of inserting into the leading strand could be observed in Classes C, D, E, F and unclassified introns (including the tentative Class H), which all lack an EN domain (Figure 20, Table 1). Also, as expected, Classes CL1 and CL2, which typically contain an EN domain, were distributed roughly evenly into both strands (Figure 20, Table 1). Even though a few CL1 introns do not contain the EN domain, a separate plot including only those introns (not shown) still showed an even distribution resembling the plot for the entire CL1 class.

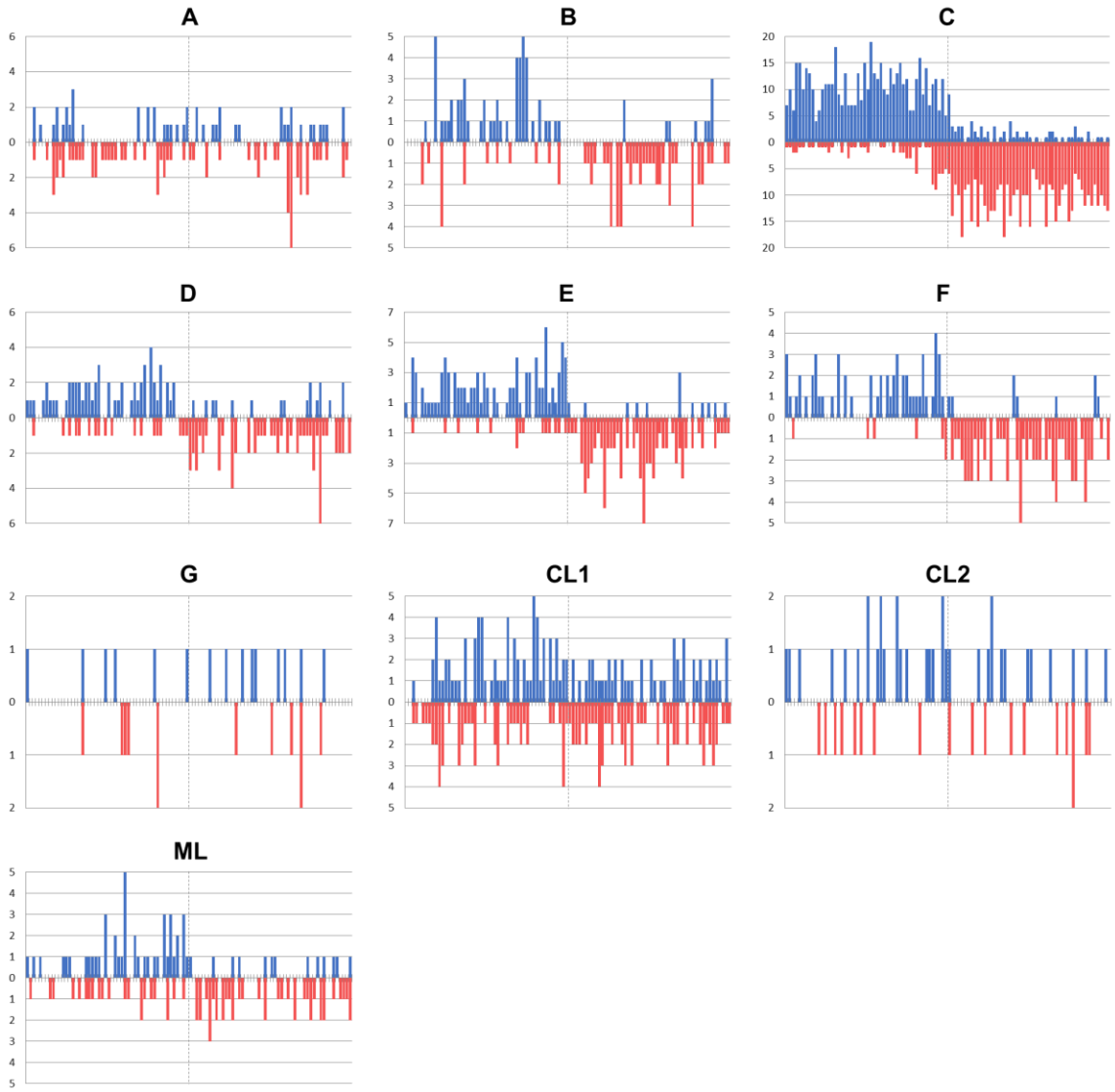


Figure 20. Numbers of introns inserted at each 1% genomic interval for different phylogenetic classes.

All genomes were divided into 100 segments while setting position 0 and 100 (x-axis) to be the origin, which was predicted based on the GC-skew. For each class, the number of introns inserted into every 1% interval were totalled for the top and bottom strands, which correspond to blue and red bars in the plot respectively. The vertical dashed lines at the 50% position indicate the expected position of the terminus. Duplicated copies between genomes have been eliminated from the dataset to prevent bias from more frequently sequenced genomes.

Table 1. Numbers of introns inferred to be inserted in leading or lagging strands.
Introns are grouped into four sections based on their genomic locations. The ratio of introns inserting into the leading and lagging strands is approximated. Whether an EN domain is present in the IEP for each class is indicated in the last column, and data used for calculation is the same as for plots in Figure 20. According to the hypothesis that EN-lacking introns prefer to insert into the leading strand (explained in the main text), an even ratio between the leading and lagging strands would be expected for EN-containing classes, while an uneven ratio (leading >> lagging) would be expected for EN-lacking classes.

	Top 0-50% (Leading)	Top 50-100% (Lagging)	Bottom 0-50% (Lagging)	Bottom 50-100% (Leading)	Leading total	Lagging total	Leading: Lagging	Has EN domain ?
A	27	23	37	37	64	60	1:1	No
B	52	10	16	52	104	26	4:1	Yes
C	537	72	85	545	1082	157	7:1	No
D	59	17	18	63	122	35	7:2	No
E	97	13	15	96	193	28	7:1	No
F	56	9	7	82	138	16	9:1	No
G	6	9	6	6	12	15	1:1	No
CL1	75	53	52	67	142	105	1:1	Yes
CL2	22	13	8	11	33	21	3:2	Yes
ML	44	13	22	42	86	35	5:2	Yes

In contrast, there were also unexpected observations. As classes that do not have an EN domain, Classes A and G showed an even distribution into both strands, while EN-containing Classes B and ML showed a bias in targeting the leading strand more frequently (Figure 20, Table 1). One explanation for this could be the genes into which these introns insert. It could be difficult or even impossible to discriminate between insertions into the leading or lagging strand and cases where the transposon has subsequently moved. As mentioned in the previous section, Classes A and G mainly insert into mobile transposons, and thus the observed distribution of these introns onto the two DNA strands could reflect the preference of the host transposons instead of the preferences of the introns. Similarly, Class B and ML introns often target housekeeping genes, and if the locations of host genes are biased in the different strands, the distribution of introns inserting into these genes would also appear to have a bias. Since the genomic locations of housekeeping genes do not change as transposons do, this could be the main reason that caused the bias observed for B and ML.

This analysis was limited by the number of host genomes that have been completely sequenced, as well as whether the complete genome shows a clear GC-skew, since this

was used for origin prediction. About 20% of the complete genomes analysed did not show a clear GC-skew and were excluded. As an alternative attempt, these genomes were submitted to the DoriC database (http://tubic.tju.edu.cn/doric/) [268]. However, no informative results could be obtained for most genomes submitted. The most common case was the DoriC-based prediction included multiple candidate origins that are far away from each other. Therefore, even though one of them was in consist with the GC-based prediction, it would be difficult to determine which of the suggested candidate could be used for genomes without a clear GC-skew. As a result, genomes without a GC-skew remained unused. However, it would still be of interest to investigate whether introns belong to such genomes, such as the Firmicutes phylum that shows biased nucleotide usage in the coding region [269], could reveal a different pattern in the insertion strand; and if so, whether this is also associated to the pattern of the homing site such introns. Although there are other methods for origin prediction (discussed in [270]), they used in this analysis due to the lack of available applications, but can be considered to be tested in the future.

Non-standard intron organisations

A standard intron organisation consists of a ribozyme RNA with the IEP residing in DIV. Almost all functional introns in bacteria are of this form and they can be easily identified from genomic sequences. Introns can also appear in non-standard forms, such as two introns directly next to each other (tandem), one intron within another (twintron), an intron that does not possess the IEP portion (ORF-less) or even more complicated nested organisations.

Such non-standard introns are generally harder to detect automatically. First, the pipeline program begins with the search for the IEP, and thus ORF-less introns are not addressed. Second, in a twintron organisation, even if the pipeline program can detect the inner intron, the outer intron would be most likely marked incomplete due to the interruption caused by the inner intron. Third, although in tandem intron organisations, each individual intron copy is expected to be detected automatically, they would not be identified as tandem to each other because the pipeline was not programmed to record their genomic locations and to examine the proximity to other introns. Other intron complexes would also be dismissed by the pipeline due to similar reasons, such as lacking the IEP or having interrupted segments.

Although attempts were made in the pipeline program for finding twintrons and tandem introns, it only detected a small number (~5) during in the 2013 update. During the 2017 update, intron complexes were thus identified manually through visual inspection of existing group II-related IEP and RNA segments (Figure 15, Methods). Even though this was not a systematic approach, it indeed revealed a rather large set of ~60 unique examples and showed a great variation in non-standard introns.

Tandem introns

Tandem intron units consist of two introns adjacent to each other without any extra nucleotides in between (Figure 15A). Each intron copy is often complete and can be individually detected by the pipeline. A total of 17 unique tandem intron units were discovered, which were divided into two major types, I and II, based on whether the intron copies are of the same prototype. Each type can further be divided into two subtypes, "a" and "b", based on the order of each intron being inserted into the chromosomal DNA (Table 2).

Table 2. Tandem intron classification.

Tandem introns are grouped based on the identity of the two copies. The insertion pattern is based on their predicted EBS-IBS interactions. Numbers listed in the table do not include multiple copies. Detailed information for each tandem unit and its copy number are provided in Appendix B3.

Main type and description	Subtype and description	In coding exon sequence	Intergenic	Total #
I. All copies belong to the same prototype	a. Downstream intron inserts earlier than the upstream intron	8	6	14
II. All copies belong to different prototypes	a. Downstream intron inserts earlier than the upstream intron	0	1	1
	b. Upstream intron inserts earlier than the downstream intron	0	2	2

The prediction of the order of intron insertion was through the EBS-IBS pairings. In subtype "a", the downstream intron is believed to insert into the genomic DNA first, followed by the upstream intron inserting at its 5' end. As a result, both the upstream and downstream introns recognise the same home site sequence, while the downstream

intron (which inserted earlier) is separated from its EBS sequence by the upstream intron (which inserted later) (Figure 21A-C). This is the most common type among all identified tandem units (15 of 17, Table 2). They are expected to be functional and can splice as a single unit, not only because many of them (8 of 17, Table 2) are found in coding sequences, but also because some have been found multiple times within the same flanking sequences (Appendix B3). A special example of this category consists of three identical introns (Du.tl.I1) in tandem with all recognising the same homing site (Figure 21B).

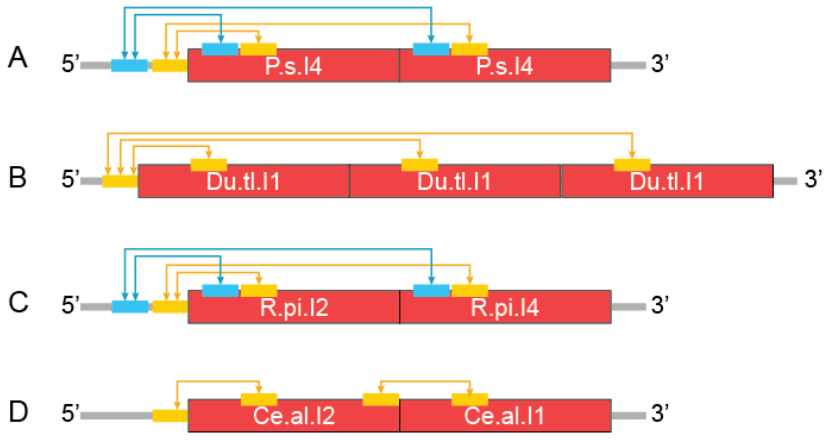


Figure 21. Examples of tandem intron units.

Introns are depicted as red blocks and given intron names are selected real examples. EBS1-IBS1 and EBS2-IBS2 interactions are denoted by orange and blue arrowed lines respectively. **A)** Type I-a. Two P.s.I4 copies in tandem with both recognising the same IBS1 and IBS2. **B)** Type I-a. Three Du.tl.I1 copies in tandem with all recognising the same IBS1. **C)** Type II-a. Different introns R.pi.I2 and R.pi.I3 in tandem with both recognising the same IBS1 and IBS2. **D)** Type II-b. Different introns Ce.al.I2 and Ce.al.I1 in tandem and they recognise different IBS1 sequences. Names of these types correspond to Table 2 and Appendix B3. Based on the EBS-IBS interactions, the downstream intron is likely to insert earlier than the upstream intron in A), B) and C), while the upstream intron is likely to insert earlier than the downstream intron in D) (see main text).

In subtype "b", the upstream intron inserted into the genomic DNA first, followed by the downstream intron inserting at the 3' end. Therefore, the upstream intron (which inserted earlier) and its IBS sequence would remain adjacent, while the 3' end of the upstream intron also serves as the IBS sequence for the downstream intron (which inserted later) (Figure 21D). Of the 17 unique tandem units, only two belong to this type, and at present, none of the two had support of being mobile as a tandem unit due to the lack of

multiple copies (Table 2, Appendix B3).

Twintrons

A typical twintron consists of two introns with one residing within the other (Figure 15B, C). The pipeline was only capable of precisely detecting the inner copy because of its intact organisation. When a twintron interrupts a gene, both intron copies have to be spliced out to maintain the gene function. This would require the inner intron first splicing out from the outer intron, followed by the outer intron splicing out from the flanking sequences. There are 25 new unique twintrons identified in this compilation, raising the total number of unique twintrons to 42. Different from tandem units, twintrons showed a larger diversity, and resulted in more types (A to F) based on the properties of both copies (Table 3, Figure 22, Appendix B3).

Table 3. Twintron classification.

Twintrons are divided based on the insertion site of the inner copy, and then by other features such as the presence of the IEP or the orientation of each copy. Numbers given in the table are only of the unique twintrons, and a full list of twintrons including duplicated copies is available in Appendix B3.

Insertion site of the inner intron	Type and description	In coding exon sequence	Intergenic	Total #
Within the IEP or DIV of the outer intron	A. Both copies have an IEP	7	12	19
	B. Inner portion contains two introns in tandem	0	2	2
	C. The inner copy is ORF-less	1	10	11
	D. Both copies are ORF-less	0	1	1
	E. Two copies are in opposite orientation	0	2	2
Within the ribozyme portion (DI, II, III, V, VI) of the outer intron	F. The inner copy resides within the ribozyme outside of both IEP and DIV	0	7	7

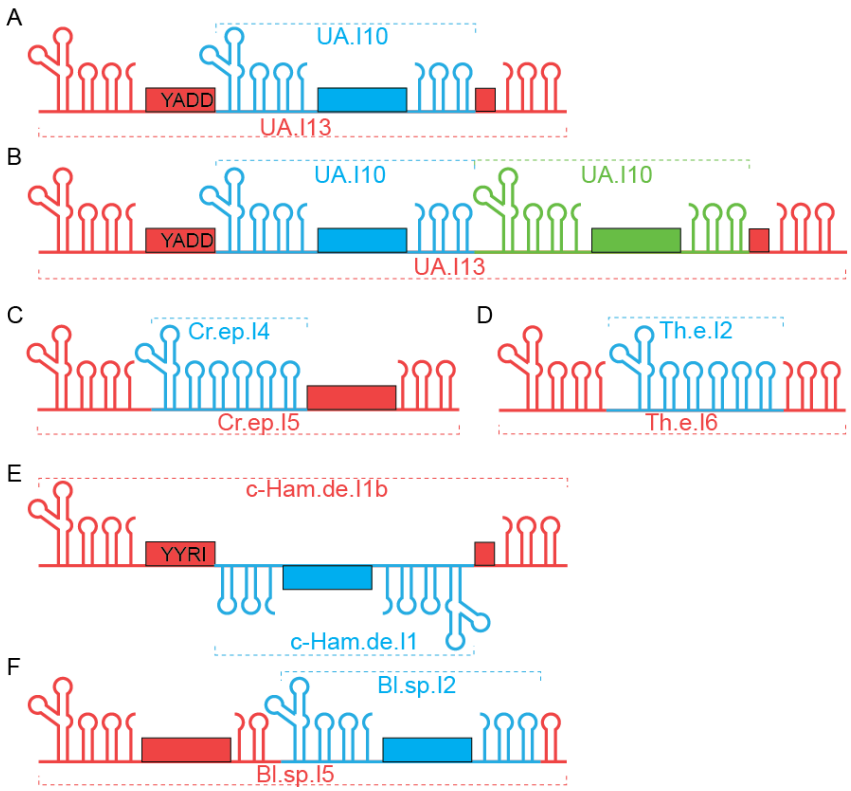


Figure 22. Examples of various twintron organisations.

Introns are depicted by a combination of stem-loops and blocks that correspond to the RNA domains and IEP portions respectively. When the inner intron interrupts the IEP of the outer, the last four amino acids at the upstream of the insertion site are specified: "YADD" is conserved in motif 5 of the RT domain, and "YYRI" is conserved in the X domain. Colours are used to differentiate multiple intron copies, and dashed lines above or below draw the boundaries of each complete intron together with the specific intron name. Figure is not drawn to scale. **A)** Both the inner and outer copies contain an IEP. **B)** The inner portion contain two IEP-containing introns in tandem. **C)** An ORF-LESS intron is inside of an IEP-containing intron. **D)** Both inner and outer introns are ORF-less. **E)** The inner and outer introns are in the opposite orientation. **F)** The inner intron resides within the ribozyme sequence in a location other than the IEP or DIV of the outer intron.

Type A has the most common twintron form, in which both the inner and outer introns contain an IEP, and the inner one resides within the IEP of the outer (Figure 22A). Type A contained 19 examples, and about one third (7 of 19) of type A twintrons reside in genes (Table 3). If considering multiple copies, about half (32 of 63) of all identified twintrons belong to type A (Appendix B3). Type A twintrons are found in various species and involve introns from almost all classes except for Class A (Appendix B3). Because of the frequent observation of multiple occurrences and the widespread distribution, type A twintrons are generally expected to be active, and a few examples have already been

experimentally confirmed [271].

Type B twintrons can be considered as a special case of type A. They contain two introns in tandem as the inner portion while the IEP of the outer intron is interrupted (Figure 22B). Two unique type B twintrons have been identified (Table 3), and one of them was also found with two copies (Appendix B3), indicating this unit is still likely to be active.

For both type A and type B twintrons, the inner introns are often found to insert immediately downstream of the essential "YADD" motif of the RT domain (Figure 22A, B). The same insertion pattern was also observed in a previous study [98], which indicates that many introns prefer to home into conserved sequences, and when the homing site is a conserved RT motif of the IEP, it would cause introns homing into other introns followed by the formation of twintrons.

In addition to twintrons that only contain standard intron copies, one or both introns of the twintron unit can be ORF-less. Type C twintrons contain an ORF-less inner intron (Figure 22C), and type D twintrons have both the inner and outer copies being ORF-less (Figure 22D). However, twintrons with an ORF-containing inner copy and an ORF-less outer copy have not been identified yet. Both type C and type D twintrons are expected to require an IEP encoded by a different intron in the genome to assist their splicing. After the inner intron splices out, the outer intron of type C is capable of expressing its own functional IEP, while the outer intron of type D would require the assistance of another IEP, which may or may not be the same IEP used by the inner intron. Type C is the second largest group in size (11 of 42) among all types, and one was found residing in a gene (Table 3). One of the 11 type C twintrons was found with four identical copies (Appendix B3), implying a possibility of this unit being transferred across individuals and could be active. In contrast, type D is the smallest group since it only contains one example with no additional copies at all (Table 3, Appendix B3), and thus it remains unclear whether this type D twintron is still active or is a remnant when both copies lost their IEPs.

Type E has the inner and outer introns residing in opposite orientations (Figure 22E). Currently, two examples have been found (Appendix B3): one contains an outer ORF-less intron and a standard inner intron, and the other contains two standard introns and has three copies in the same genome, which suggests the latter example is still active.

For the latter example, the inner intron inserts after sequence "YYRI", which belongs to a conserved region in domain X. Nonetheless, since the inner intron targets the sequence on the other orientation, which does not appear to code for a meaningful ORF, it is not yet clear whether the "YYRI" region could be a second insertion site favoured by these inner introns, or the inner introns only target the specific homing sequence, regardless whether it is associated to the IEP on the opposite strand.

Type F is different from all other types in terms of the location of the inner intron, which is within the ribozyme but not in the DIV nor the IEP. Of the seven type F twintrons, almost all inner introns were found to target the linker sequence between DV and DVI (Figure 22F), and one of them contains three copies that indicate its splicing ability (Appendix B3). The outer introns of type F are from different classes (C, D, E, unclassified, Appendix B3) and there is no sequence similarity in their linker sequences. Therefore, in contrast to types A and B that were formed by introns homing into conserved IEP sequences, type F twintrons could be formed when introns home into conserved RNA structures (e.g. DV).

Overall, twintrons have shown a great variety in their organisation, although having two standard IEP-containing introns is the most common form of presently identified twintrons. Most types are expected to be capable of splicing in an order of inner-to-outer based the copy number of some examples (Appendix B3), but there is also a possibility that some (if not all) multiple copies were caused by genomic duplication events, which can be verified by comparing both the intron and flanking sequences. As the next step, the large variety of different twintron structures can be further investigate, and the examination of the sequences in their neighbourhoods should be considered.

ORF-less introns

ORF-less introns lack the IEP portion entirely or contain only a short piece of remnant of the IEP. They are expected to require the assistance of other IEPs *in trans* in order to splice or be mobile. They could not be detected by the pipeline, which was designed to search for introns relying on the IEP similarity and domain completeness. Earlier searches identified many ORF-less examples by beginning the search for the conserved domain V of the RNA ribozyme [272]. This time, such examples were directly detected through visual inspection this time by looking for closely located 5'- and 3' RNA

segments where the IEP is absent (Figure 15C). ORF-less RNAs can be folded as standard introns, and traces of IEP are often observed, ORF-less introns are thought to have originated from loss of the IEP. Very often, ORF-less introns are involved in twintrons or other complexes (next section).

Irregular intron organisations

During the step of finding tandem introns, twintrons and ORF-less introns, it was noticed that some introns could form even more complicated chimeric organisations. As observed most commonly, they involve segments from different introns being mixed with each other, such that no complete individual intron could be identified. These are believed to be defective and thus marked as intron graveyards. Interestingly, there are also cases when all segments can be reassembled back into individual introns, and they may still be functional if each unit splices out in order. This section gives a few examples of such complexes (Figure 23).

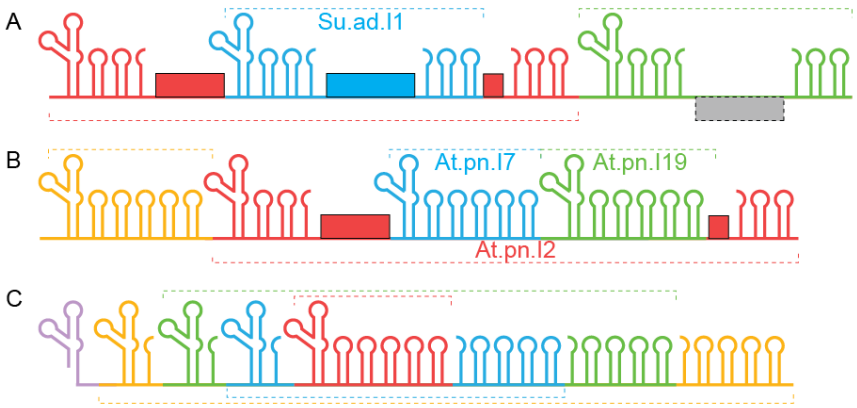


Figure 23. Example of non-standard intron complexes.
Introns are depicted using the same style as in Figure 22. A grey block corresponds to a transposase as shown in panel A. Some introns are not yet finalised and thus their names are not provided. Examples are all from genomes CP003180 and FO818640.

A simple example from genome CP013008 (*Arthrospira platensis YZ*) shows a twintron in tandem with another intron that contains a transposase instead of a group II IEP encoded on the opposite strand (Figure 23A). In genome FO818640 (*Arthrospira sp. Str. PCC 8005*), there is a complex consisting of an ORF-less intron in tandem with a twintron, which consists of two ORF-less introns in tandem as the inner portion of the

twintron (Figure 23B). In the same genome, another example shows a nested twintron organisation in which the same intron repeatedly targets its own DII at the same site (Figure 23C). Interestingly, this situation was also found in a different location in the same genome, however, the two units are not exactly identical as they differ by the presence of a portion at the 5' end (Figure 23C, purple). Therefore, this unit may initially form by the same intron targeting itself over time. While it is possible to be active, the 5' portion could be lost due to the lack of the corresponding 3' end during the homing reaction. Alternatively, multiple copies could be generated through genome duplication, but this is less convincing by comparing their flanking sequences.

In addition to these examples, genomes CP013008, FO818640 and CP000393 (*Trichodesmium erythraeum IMS101*) were found to be extremely rich in such examples, and also contain a large number of inactivated intron graveyards (Appendix B4), indicating these genomes have undergone frequent genomic rearrangements.

Updates to the database for group II introns

As introduced earlier, two updates performed in 2013 and 2017 to predict group II introns from the GenBank database have greatly increased the total number of intron prototypes, the number of multiple copies, and also some non-standard intron forms. However, during the first attempt of appending all this new information into our group II intron database, it was noticed that our old database was designed to use one single table to store all intron data. Since our previous database only had one record for each intron prototype, using a single-table would not often cause problems in performance. However, the single-table limits the flexibility for data manipulation and causes a large amount of redundancy, especially when multiple copies of each prototype are added. For example, to store one intron prototype plus nine of its multiple copies into the single table, all basic information (e.g. name and class) must be repeatedly recorded for each copy, resulting in nine of them being redundant. Also, when any of such information is to be modified, for example, to change the class name from "bacterial A" to "bacterial B", the same procedure of editing must be repeated for all 10 copies. Currently, this process is not automated by our dataset and could only be done manually. Not only does this process increase the amount of manual work, it would also be more prone to human errors that would cause data inconsistency.

Therefore, I performed updates to our group II intron database, including a redesigned

database structure and new scripts for the website to provide new features. All these updates have been described in detail in a separate document as a user manual for the lab, and this section will only give an overview about the new database design.

As depicted in Figure 24, the new database structure was designed to use multiple tables for different types of data. Each table contains two or more "fields", in which one of them is set to the "primary key" (also referred as "ID" in this section), which serves as a unique identifier for each record stored in the table. Different tables are therefore connected through the primary keys (Figure 24).

Two main tables, named "prototype" and "intronseq" respectively, are used to store the majority of intron data (Figure 24 [1-2]). The former is specific to intron prototypes and stores general information such as the name, class ID and host organism ID (Figure 24 [1]), while the latter is specific to each individual intron copy and stores information such as the GenBank accession and coordinates and the boundary of the IEP within the intron DNA (Figure 24 [2]). Information shared across introns, including the IEP- and RNA-based class names, the host organisms and the phylogenetic domains are stored in separate tables (Figure 24, [3-6]). In addition, tables "ge95" and "ge95_member" are used to store groups of prototypes that are >= 95% identical in sequence (Figure 24, [7-8]), and tables "tandem" and "twintron" are used to store tandem introns and twintrons (Figure 24, [9-10]). Detailed descriptions of all tables and fields are provided in Appendix B5.

The redundancy caused by using a single table is eliminated by using multiple tables, because shared information (e.g. class name) is only recorded once. Different pieces of data in multiple tables can be linked to each other, which enables a higher flexibility. Meanwhile, chances of human errors during data manipulation are thus also decreased.

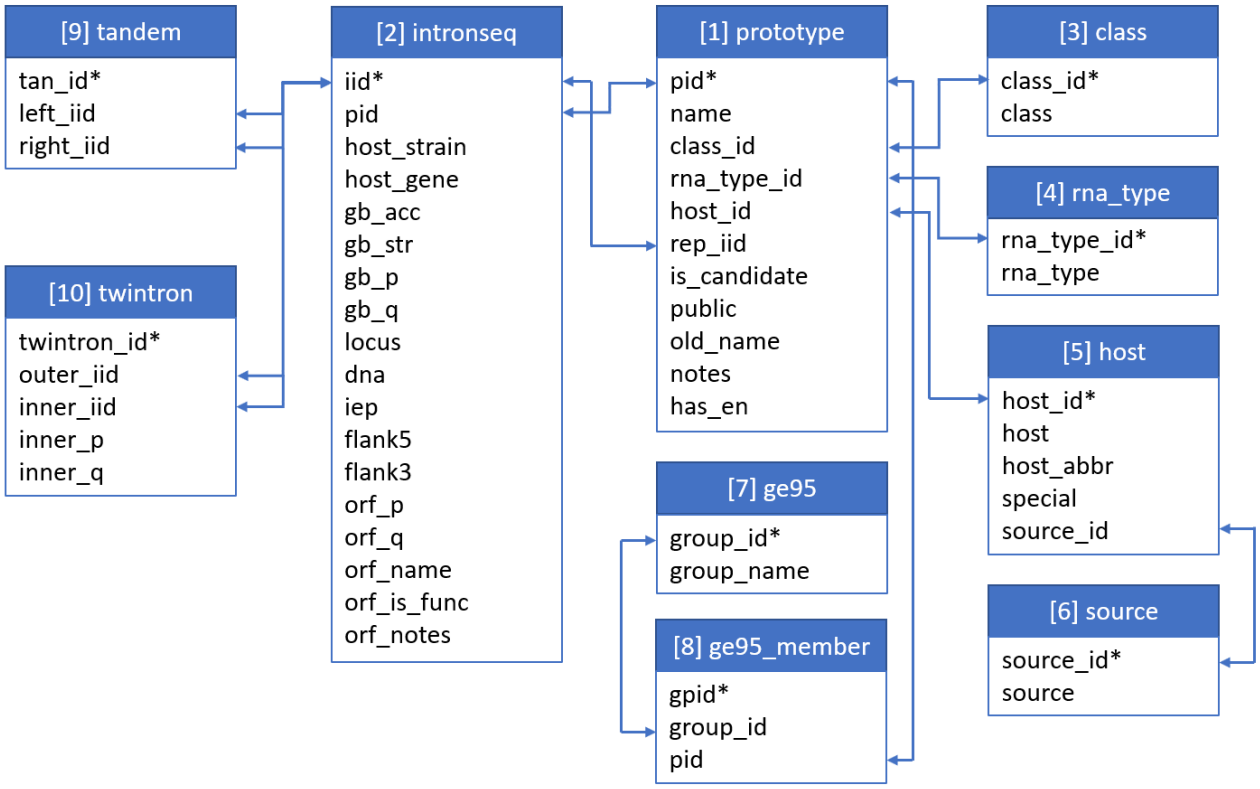


Figure 24. Data structure of the revised group II intron database.

Multiple tables are used to store different types of data. Each table is present as a box, while the name of table is given on top in the shaded area. The number surrounded by square brackets does not belong to the table name, and is only used to number the tables for referencing convenience. Names of all fields in each table is listed in the open box below the title, and detailed explanations of all fields are provided in Appendix B5. The field used as the primary key of each table is indicated by an asterisk sign ("*"). Solid arrowed lines depict the relationship between tables.

Conclusion

The project described in this chapter first collected group II introns in bacteria through an automated approach, and two updates performed in 2013 and 2017, respectively, increased our collection of unique intron prototypes by about 4-fold. Phylogenetic analyses have revealed a new Class G, a tentative new Class H, and ~15 unclassified introns. This larger dataset enabled an update of the RNA consensus secondary structure for Class A based on the new set of over 40 Class A introns, which would be a more convincing and accurate than the old consensus, which was made from only three members that were nearly identical.

All intron prototypes were searched for other identical copies, and all these data have been stored in our group II intron database, which has been technically improved to reduce data redundancy and provide a more user-friendly interface. All new introns have been integrated into current alignments for IEP and RNA. However, due to the amount of time required, they have not been completely refined and will require future corrections. As the next step, new sequence profiles can be created from these refined alignments and used to perform future searches. Because of the increase of size for each class, the RNA alignments are especially expected to provide better reference for automated intron folding in the future, reducing manual requirements. This will allow the intron database to be dated with the expanding number of sequences on GenBank with minimal amount of manual labour required (e.g. for sequence refinement and correction).

While more introns and potential new classes are expected to be discovered in the future as the GenBank database increases, the current dataset has shown a large variety in group II introns in many aspects. First, different intron lineages tend to more often target certain types of host genes during the homing reaction, including genes, transposons or intergenic areas. This was first concluded by manually examining a small intron subset, and then supported by an automated prediction based on BLASTX. The manual examination should be mostly accurate, it required a lot of manual work and was time consuming. In contrast, even though error-free results can never be produced by the automated approach, it provided an overall similar pattern compared to the manual examination. The automated approach can thus be used in the future for similar analysis, but will first require some improvements to increase the accuracy for protein identification. Possible suggestions would focus on more hits instead of only the top hit, and the program should also be able to judge between the "top" hit and the "best" hit,

which would usually be different if the intron reside at either end of a protein sequence. In addition, the host organism of each intron appears to correlate well with the IEP-based phylogeny at the phylum level, which indicates introns of the same class are actively transferring between similar species, while occasional horizontal transfer events occur between more distantly related species, thus gradually introducing a different class of intron into a new group of hosts (phyla).

Second, most classes that lack the EN domain in the IEP showed a bias in inserting into the leading strand, which agrees with the hypothesis that such introns tend to use the lagging strands of the DNA replicate fork to initiate DNA synthesis of the other strand of insertion. However, this analysis only considered genomes with a clear GC-skew, which was then used to predict the origin, but it was also noticed that many genomes that were excluded in this analysis also contained multiple introns. It would be of interest to investigate further the preference of insertion strand for these introns if the origins of such genomes can be predicted, and whether they could affect the current conclusions.

Non-standard intron organisations, including tandem introns, twintrons, ORF-less introns and more complicated forms, were discovered from the updated database. They were often missed by the pipeline program either because they appeared to be incomplete intron segments or lacked the IEP portion so that they could not be detected by the pipeline. Through visual inspection, over 50 such non-standard examples were found and analysed. They are expected to be active if all introns splice out sequentially. Many such complexes have multiple copies in the same genome, which also implies they are, collectively, still active mobile elements. This project has provided a procedure with several scripts that are ready to use. In the future, the visual inspection-based strategy is expected to be programmed to identify automatically introns and provide suggestions of potential intron complexes for manual proofreading, and scripts used for visualising the organisation of intron complexes can be improved and integrated into our pipeline to generate user-friendly graphics.

Overall, this study provided an updated collection of group II introns and is should cover most typical introns as of spring 2017. In addition to continuous identification of more standard introns as the GenBank database increases, future searches will focus on non-standard introns as well as novel introns, which may either have DNA sequences less similar to all currently known introns but still fold into a conserved secondary structure, or can be sequentially detected but have other secondary structural features. Novel introns

are expected to form new classes, and could give new insights into the diversity of group II introns and their phylogenetic relationships.

Chapter IV. Phylogenetic analysis of group II introns

Abstract

Many phylogenetic analyses were performed on an enlarged set of group II introns in order to depict their evolutionary history. Phylogenetic trees were separately constructed using sequences from the IEP and RNA, and were constructed on different scales: the full that included all introns (global set), and individual subsets for each class of introns (class-specific sets). The relationships between classes were consistent in different trees, but all lacked support. Among trees constructed using the same data type (IEP or RNA), class-specific trees generally provided a higher resolution compared to the global trees, but there were no significant topological differences found through manual comparisons. Similarly, there were no major differences found between trees constructed using different data types, agreeing with the hypothesis that the IEP and RNA were coevolving. Although attempts were made to obtain more formal evaluations between these trees through the SH topology test, it was finally concluded that trees of group II introns may not be suitable for topology tests as minor yet common topological disagreements would often cause the null hypothesis to be rejected. In addition, the larger dataset did not appear to be sensitive to taxon-sampling but seemed to be affected by biased sampling that was based on the GC-content, indicating some topological differences observed between this and the previous studies [181] could be due the use of different datasets. Additional trees constructed using the morphologic data showed consistent topology between classes, which provided evidence to the unsupported relationships between classes. Finally, by including many non-group II RT sequences, the oldest group II intron class was shown to be most likely Class C, which is the smallest in both IEP and RNA portions and agreed with a parsimonious evolutionary history.

Introduction

Group II introns are large catalytic RNA containing an intron-encoded protein (IEP) to assists their splicing and mobility reactions *in vivo*. The RNA portion shows little sequence similarity between individuals in general but they can all fold into a conserved secondary structure of six domains (DI-DVI) surrounding a central wheel (Figure 25A) [102, 103, 104]. Domain I is the largest domain and it is responsible for recognising the 5'- and 3' exons through exon-binding sites (EBS) 1, 2 and 3. It serves as a scaffold and contains many tertiary interactions that are essential for the RNA to fold into the correct form [273]. Domain II helps to stabilise the RNA structure but is less conserved for the splicing reaction [274, 275]. Domain III interacts with domain V and enhances the catalytic reaction [276, 277]. Domain IV contains the ORF of the IEP, but this domain itself is not required for the splicing reaction. Domain V is the most essential domain for splicing by forming the catalytic core [103, 104]. Domain VI contains a bulged adenosine near the 3' end, which initiates splicing by providing its 2'-OH at the beginning of the first transesterification step (Figure 6) [18, 278, 279, 280]. Based on tertiary base pairings and structural features, the RNA can be divided into three major types, IIA, IIB and IIC [18, 103, 104, 278, 281].

In contrast, the IEP of different introns are more conserved in sequence, and a typical IEP consists of multiple domains: RT, X, D and sometimes EN. The RT domain contains conserved motifs 0-7 and has similar sequence with other RTs (Figure 25B). Based on the IEP sequence, introns can be divided into classes named bacterial A to G, chloroplast-like (CL) 1 and 2, and mitochondrial-like (ML) [106, 272, 282]. It has been noticed that each IEP-based class corresponds to a certain RNA type: ML introns have IIA RNA structure, bacterial C introns have IIC structures, bacterial A is a hybrid between IIA and IIB, while the rest all have IIB structures [272, 281]. Together with that the IEP binds to the RNA ribozyme during both splicing and mobility reactions *in vivo*, the intron IEP and RNA are hypothesized to have coevolved [272, 281, 283].

Previously, a phylogenetic study performed in 2009 [181] (which will be referred to "the 2009 study") used a large number of group II introns from all identified classes and constructed trees using both IEP and RNA sequences, and trees were compared to test the hypothesis of their coevolution. That study obtained robust clades for each defined class and supported the hypothesized coevolution between IEP and RNA as their sequences yielded agreeing topologies in general [181]. However, the 2009 study had

many limitations and unsolved questions. The first problem was that no support was obtained beyond the class level, and thus the relationships between classes remained ambiguous. The second was a potential conflict between the IEP and RNA found for the two CL classes: the IEP-based tree showed that each of the two classes clearly split into two strong subclades (CL1A, CL1B, CL2A, CL2B), while CL2 was internal to CL1; but in the RNA-based tree, CL1 and CL2 appeared as sister clades and no subclades could be observed.

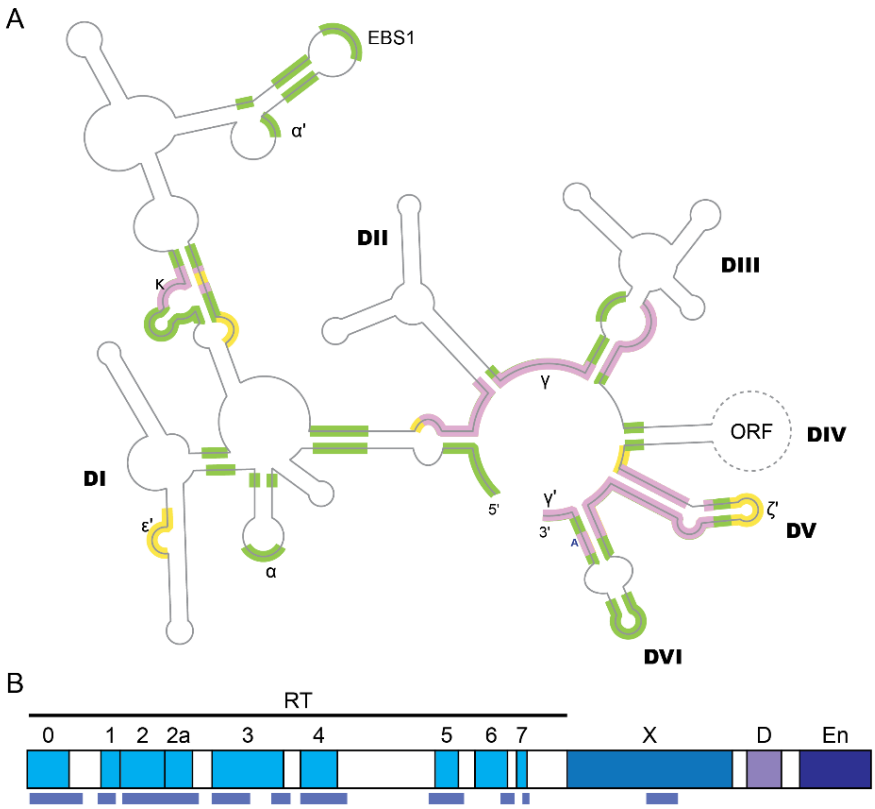


Figure 25. Typical group II intron RNA and IEP structures.

A) The secondary structure of the ribozyme RNA with the ORF of IEP in DIV being truncated and represented by a dashed circle. The green and yellow shadings are specific to regions covered by the global lenient RNA mask and the alt-RNA mask respectively, while pink shadings indicate overlapped areas between the two masks. **B)** The secondary structure of the IEP including domains RT, X, D and EN as well as all RT motifs 0-7. Thick blue bars underneath indicate areas covered by the global lenient IEP mask.

In addition, the 2009 study performed topology tests, which indicated the IEP- and RNA-based trees were conflicting. Although the reason could not be identified, two possible explanations were suggested: 1) the observed inconsistency reflects actual evolutionary difference, or 2) environmental and functional constraints could cause sequence convergence during evolution, and thus some areas may appear to be similar in sequence even from less related classes, resulting them to be incorrectly placed close to each other [284]. A larger dataset was therefore suggested by the 2009 study. Since sequences from the same class are more similar to each other compared to those from other classes, class-specific trees should have a higher resolution. However, the 2009 study was limited to only about 70 introns, and class-specific trees could not be tested that time.

The 2009 study also indicated trees suffered from sensitivities to taxon sampling and base composition heterogeneity, which can be reflected by the GC content [284]. While the former would produce different topologies from similar datasets, the latter would group distantly related taxa together if they have similar GC contents. Nonetheless, it could not be further tested previously by dividing the total dataset into more subsets, due to the limitation of a small dataset.

As described in Chapter III, two updates increased the total number of introns stored in our database from about 300 to about 1200. Therefore, by taking advantage of the enlarged dataset, the project described in this chapter aimed to solve questions left from the 2009 study. Not only this project constructed a series of trees using both the IEP- and RNA-sequences, it also tested the application of the paired-site model "RNA7A" [285] to investigate the effect of considering secondary structural information of the group II ribozyme as suggested by the 2009 study. Trees were inferred using introns from all classes (global dataset) or from only one class (class-specific dataset), and tree topologies were compared in order to identify potential conflicts between the IEP and RNA. Trees were subjected to the SH topology tests, and also tested for the sensitivities to taxon sampling and base composition heterogeneity. This project also extended the previous study by constructing trees using the morphological data in order to restore information dismissed due to the lack of sequence similarity. Finally, other types of RTs were incorporated with group II introns in order to investigate the possible origin of group II introns.

Methods and materials

Data collection, sequence alignment and masks

This project used 661 group II introns from bacteria and archaea (Appendix D1), which were collected as described in Chapter III. All introns are predicted to have a functional IEP and a normal ribozyme RNA that can be folded into six domains. These introns belong to 10 classes, including bacterial A to G, CL1 and 2 (chloroplast-like) and ML (mitochondrion-like). This dataset also included six unclassified individuals: B.ce.I2, Hn.pr.I2, Ho.hb.I2, Su.tt.I2, U.a.I4 and Hp.au.I3. All IEP sequences were automatically aligned by program HMMER [214] using a group II RT profile established previously by our lab, followed by manual refinements. RNA sequences excluding DIV were manually aligned based on their structural features.

The 661 introns were divided into subsets based on the class in order to create class-specific trees (herein, referred as "class-specific"), also, all introns were used at the same time to create trees including all classes (herein referred as the "global"). For both global and class-specific datasets, several sequence masks were applied to each alignment to specify more conserved sites. The IEP masks were made based on sequence, while the RNA masks were based on both sequence and structure. Generally, three IEP masks were created, and from the longest to shortest, the masks were named "lenient", "medium" and "strict" respectively (Table 4). Two masks, "lenient" and "strict", were generally created for RNA alignments, while the "medium" mask was not created for the RNA (Table 4). Of note, the global lenient IEP mask (the longest mask for the global IEP dataset) was the same as the sequence mask used in the 2009 study [181].

For the RNA dataset, an additional mask, named "no gap", was derived from the "strict" RNA mask by removing all sites that contained at least one missing characters (gaps). In order to maintain the structural information, if the deleted site is involved in a base pair, both sites were excluded from the mask. All masks used in this project along with their sizes are listed in Table 4. As will be described later, another global mask, named "alt-RNA" (Figure 25), was also used for the global RNA alignments only during tests for taxon sampling (see later). This mask covered the same area as used in the 2009 study [181], and was created based on only sequence.

Table 4. Sizes of masks used in the phylogenetic analysis.

A) Sizes of masks of global alignments. **B)** Sizes of masks of class-specific alignments. Depending on the type of data, the unit is amino acid (aa) for IEP and nucleotide (nt) for RNA and RNA-STR. "N/A" indicates the corresponding mask was not created.

A. Size of global masks.		
	IEP (aa)	RNA (nt)
global lenient	285	206
global medium	245	N/A
global strict	168	170
global alt-RNA	N/A	138

B. Size of class-specific masks.						
Class	IEP (aa)			RNA (nt)		
	class lenient	class medium	class strict	class lenient	class strict	class no gap
A	502	N/A	419	673	633	619
B	499	388	316	573	486	401
C	394	328	248	380	367	303
D	395	335	272	482	N/A	447
E	411	340	280	455	412	349
F	354	316	264	492	475	403
G	375	334	266	582	561	519
CL1	421	374	339	547	500	328
CL2	476	376	347	524	469	355
ML	465	347	299	530	488	349

Model test and phylogenetic inference

Optimum models for the IEP and RNA datasets were selected by ProtTest [286] and jModelTest [287] respectively using default settings for each program. Based on evaluations using both Akaike information criterion (AIC) and Bayesian information criterion (BIC), the LG+I+G model for IEP and the GTR+I+G model for RNA were preferred by the majority of alignments and were used in this study, while "GTR" stands for "general time reversible", "G" stands for the gamma model of rate heterogeneity, and "I" stands for the estimate of the proportion of invariable sites. In addition, the paired-site model "RNA7A" [285] for RNA alignments with structural information was also used for the RNA dataset, and trees generated under this model will be referred as "RNA-STR" in the following text.

Unless otherwise indicated, trees constructed in this project were all done by program RAxML [226] with 1000 bootstrap replicates using models described above (command written as "-m PROTGAMMAIRTREV" while running RAxML). A few trees were

constructed using the program MrBayes [228], in which the model for amino acids was specified as "mixed" instead of LG. For MrBayes, the minimal number of generations was set to 40 million, or until the average standard deviation of split frequencies (ASDSF) was lower than 0.01 as suggested by the user manual, whichever came first. Four chains were executed for each run, and the sampling frequency and diagnosing frequency were both set to 5000. All programs were running on the Bugaboo server at WestGrid (http://www.westgrid.ca).

Taxon sampling analyses

Two types of sampling were performed. The first type selected four subsets that were evenly sampled along the global lenient IEP tree. Each subset had ~140 introns in total, which included ~20% from each class and all six unclassified introns (Appendix D1). The second type was based on the GC content and had only one subset, which included 196 introns with a GC content of 47.1 ± 5% (43-52% inclusive) (Appendix D1), while 47.1% was average GC content among all 661 IEPs. Only the lenient IEP, lenient RNA and the alt-RNA masks were used to construct these taxon sampled trees, because only minor differences were observed between various masks (see Results). Trees were constructed by both RAxML and MrBayes. To reduce the computational time, only 100 bootstrap replicates were requested for RAxML.

Shimodaira-Hasegawa topology test

The SH test was performed at both global and class levels using program CONSEL [288]. Site-by-site likelihoods were calculated by "codeml" from the PAML package [289] using the LG model (for amino acids) and "baseml" using the GTR model (for nucleotides). In order to reduce computational time, the SH tests of global trees only used the four sets of taxon sampled trees, which each set contained three global trees constructed by the lenient IEP, lenient RNA and alt-RNA masks. During the SH tests for class-specific trees, seven IEP-based trees (three subtrees extracted from global lenient, medium and strict trees, three class-specific lenient, medium and strict trees, and one class-specific tree using the global strict mask) and 10 RNA trees (two subtrees extracted from global lenient and strict trees, six class-specific lenient, strict and no gap trees using the GTR and RNA7A models) were used.

Morphological tree

Morphological data were collected from both IEP and RNA alignments. Of the IEP, they included the presence of EN domain, the average length of each RT motif, X and EN domains, as well as the degree of sequence similarity across classes (Appendix D2). The degree of sequence similarity was evaluated using HMMER-based comparisons. First, three class-specific profiles were created for various portions of the IEP alignment: areas before RT motif 0, the entire domain X, and the entire domain EN if present. Next, each profile was compared to the corresponding portion for all intron individuals by HMMER, and the average E value of introns from the same class was calculated, and used to evaluate the degree of similarity for the same portion across different classes.

Of the RNA, morphological characters were mainly based on secondary structural features, such as the length of a conserved stem, the presence of less conserved structures, and the structure formed by a specific area (e.g. whether a stem has no mismatches or contains an internal loop) (Appendix D2). Features collected from both the IEP and RNA were coded as either binary or morphology data: the former uses either the digit "0" or "1" to indicate whether a feature is absent or present, the latter uses all 10 digits (0-9) to divide each feature into up to 10 subtypes.

There were 37 and 98 morphological characters extracted from the IEP and RNA respectively. Among all characters, 49 were further chosen to form a "more important" subset, in which selected features either correspond to an essential tertiary interaction (e.g. IBS1-EBS1), or are related to the inclusion of major functional domains (e.g. the EN domain of the IEP) (Appendix D2). During the phylogenetic inference, the two data types were partitioned to use different models for the binary and morphology data, and four morphological trees were constructed by MrBayes using: 1) all 135 characters, 2) only the 37 IEP characters, 3) only the 98 RNA characters, and 4) the 49 "more important" characters.

Trees with external non-group II RTs

Several non-group II RTs were aligned to group II intron IEPs individually, including: AbiA, AbiK, AbiP2, DGR, retron, PLE, retroplasmid, RVT, TERT, G2L1-5 and unknowns 1-16, which were collected from both Chapter II and other publications (sources are specified in Appendix D1) [14, 22, 26, 33, 237, 290]. To reduce the computational time,

only 67 group II RTs was used here as representatives (Appendix D1). Program HMMER was used to align non-group II sequences using a sequence profile named "RVT1" provided by Pfam (Pfam ID: PF00078), and the alignment was integrated with the group II IEP alignment followed by manual refinements. Individual trees were constructed for each type of non-group II RTs using RAxML with the LG model and 100 bootstrap replicates.

Results and Discussion

This project used 661 group II introns and aimed to resolve a better relationship across classes using both the IEP and RNA data. This dataset was much larger compared to the previous study published in 2009 (Figure 26) [181], and included two new classes A and G and six unclassified introns (Appendix D1). In the following text, various analyses will be described, and trees will be compared with the 2009 study [181]. Although both maximum likelihood and Bayesian inference were initially planned to be used in constructing all trees, only the maximum likelihood method was used, while the Bayesian method was only used for a few analyses because it generally required a much larger amount of time and computational power.

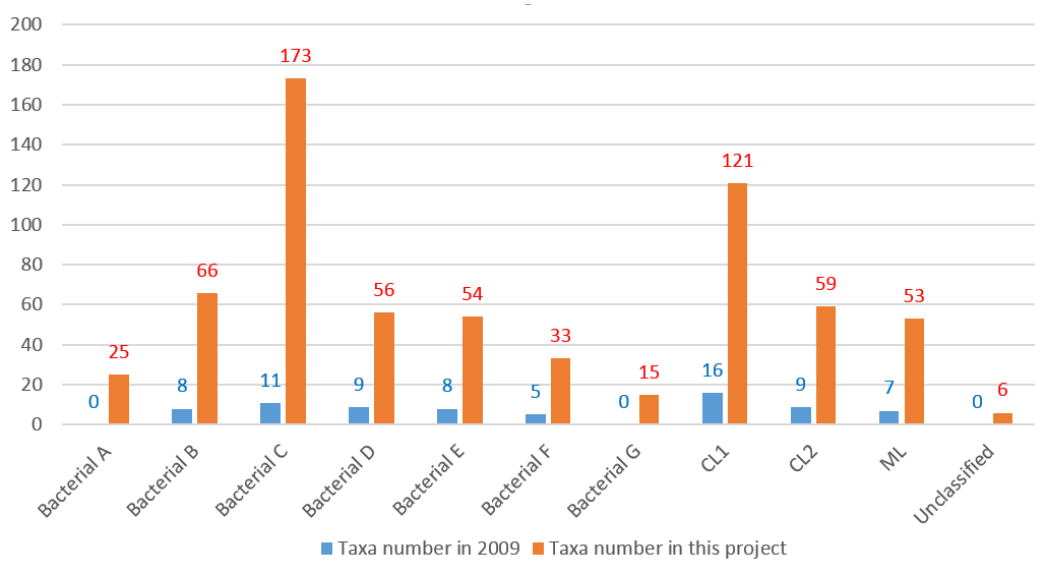


Figure 26. Increase of dataset used in this project compared with the 2009 study [181].

Among the six unclassified introns included in this study, Hp.au.I3 was noticed to locate differently from the other five (see below). Therefore, in this chapter, unless otherwise specified, "unclassified" introns will be referring five unclassified introns (B.ce.I2, Hn.pr.I2, Ho.hb.I2, Su.tt.I2, U.a.I4), while Hp.au.I3 will be referred to directly by name.

The ORF-based global phylogeny

Three global trees based on the IEP sequences were constructed by RAxML using masks "global lenient", "medium" and "strict", which covered 285, 245 and 168 aa in length respectively (Table 4A). Most classes were monophyletic and supported (bootstrap >= 75%) in at least two trees, and all three trees showed nearly identical topology in terms of the relationships between classes, although without supports (Figure 27B, C). The two new classes, A and G, were both located close to B, ML and CL in all three trees. Five out of the six unclassified introns always scattered between C and F, and the only exception was Hp.au.I3, which was close to the cluster of A+B+ML. Overall, this arrangement agrees with a later phylogenetic tree using 1275 taxa as mentioned in Chapter III (Figure 18).

It was initially expected that a larger dataset could resolve the relationship between classes. However, regardless of the consistent topology seen in all trees, no support could be observed for relationships between classes. When these trees were compared to the global IEP tree obtained in the 2009 study, which used the same global lenient IEP mask (Figure 27A) [181], topologies of the three new trees showed agreements by putting C and F together, B and ML together, and CL1 and CL2 together (Figure 27). But there was also a difference involving Class E, which was sometimes grouped with B and ML with a support in the 2009 study [181], while this time, such an arrangement was never observed in any of the new trees, which could suggest that the overall tree topology is more robust while using a larger dataset. In general, even without supports, all classes often fall into four larger groups: C+F+unclassified, D+E, A+B+ML+Hp.au.I3, and G+CL1+CL2 (Figure 27B, C).

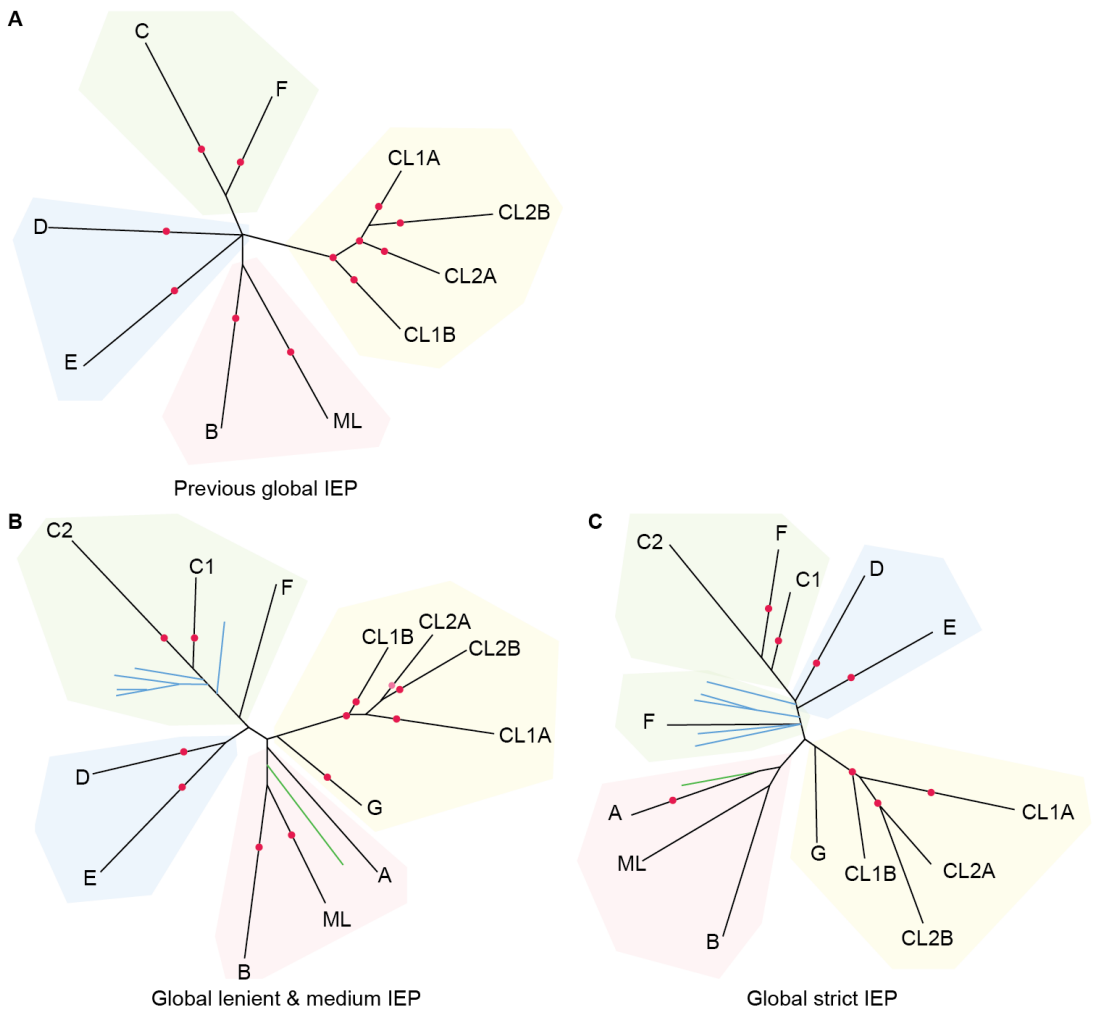


Figure 27. Global IEP trees inferred by various data types and masks.

Trees are simplified at the class or subclass level. Unclassified introns are drawn in blue except individual Hp.au.I3 that is in green. Supported nodes (posterior probability >= 0.95 for panel A or bootstrap >= 75% for B and C) are indicated by a red dot. Coloured shadings correspond to the trend as in the main text. **A)** Adapted global IEP tree from the 2009 study [181]. **B)** Global lenient and medium IEP trees, which are merged because they showed the same topology and only differed by the support level at clade CL2A, which is marked by a dot in faint red. **C)** Global strict IEP tree. Tree in A) was constructed by MrBayes and trees in B) and C) were constructed by RAxML.

There were also a few differences involving supported clades between trees constructed in this project and in the 2009 study. The first concerned Class C, which used to be a supported clade in the 2009 study (Figure 27A). This time, all Class C introns together still formed one single clade but only had low support. Instead, it further contained two supported subclades, which are herein referred to C1 and C2, corresponding to the earlier and later branched clades respectively (Figure 27B, C). However, this was not

considered a major difference. The 2009 tree only contained one intron (D.p.I1) belonging to the C1 clade [181], but it was still the earliest branching intron of Class C. Judging from results obtained during the taxon sampling test in which Class C was supported again (see below), the lack of support of Class C in the three global IEP trees should not affect the current classification, and thus both C1 and C2 subclades will still be classified as Class C instead of being reassigned to different classes.

Similar to the situation of Class C, the four subclades of CL1 and CL2 (CL1A, CL1B, CL2A, CL2B) used to be supported in the 2009 study, but this time, only CL1A was supported in all three IEP trees (Figure 27B, C). Like the decision for Class C, the lack of support for some CL subclades was considered minor because all four clades were always monophyletic and showed similar arrangements as seen in 2009, in which CL1B was always the earliest branching clade.

The next difference concerned Class F, which was never supported in any of the three IEP trees, and some F taxa were even observed being mixed with other introns from Class C or unclassified individuals in the global strict IEP tree (Figure 27C). Similar situations were observed among many IEP trees constructed in the 2009 study [181], as well as trees constructed before this project using different datasets, indicating Class F is less robust compared to other classes regardless of the size of dataset. However, Class F was still believed to be one class, not only because all Class F introns were often placed into one clade while using larger masks (lenient and medium) (Figure 27B), but also because their similarities in the RNA structure. Furthermore, this was supported by the later taxon sampling test (see "taxon sampling" section).

Across the three global IEP masks, the lenient and medium masks differed by ~40 aa characters (Table 4A) and resulted in almost identical topologies, except that the subclade CL2A was supported using the lenient mask, but lacked support using the medium (Figure 27B). The strict tree differed from the lenient mask by ~120 aa characters (Table 4A) and resulted in weaker supports in general, but still showed a similar topology compared to the other two masks, except for the polyphyletic situation of Class F as stated above.

To evaluate whether the mask size would affect the tree, it was compared to the resolution of each tree, which was used as a measure of tree confidence. The resolution of tree was calculated using the number of supported nodes divided by the total number

of nodes. As shown in Figure 28, the general trend is that larger masks usually result trees in a higher resolution. In addition, because the three trees showed no significant disagreements in the relationships between classes and because the most stringent mask (global strict) was suspected to cause problematic arrangements (e.g. class F), only the global lenient IEP mask was used in most of the following analyses.

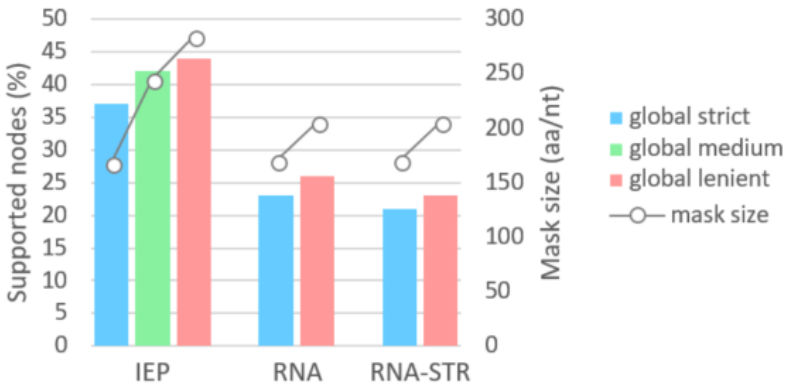


Figure 28. Percentage of supported nodes and mask size of global trees.
The first y-axis (left) corresponds to the percentage of supported nodes (bootstrap >= 75%) for each global tree. The second y-axis (right) corresponds to the mask size, and the unit is aa for IEP trees and nt for RNA and RNA-STR trees.

The RNA-based global phylogeny

Two global trees based on the RNA sequences were constructed by RAxML using masks "global lenient" and "global strict", which covered 206 and 170 nt in length respectively (Table 4A). Because the group II ribozyme is known to be more conserved in structure than in sequence, the paired-site model RNA7A [285] designed for structured RNA molecules was tested in addition to the best-fit GTR model estimated by ModelTest [286, 287]. Both the RNA7A and GTR models were applied onto the same set of RNA alignments. To differentiate trees resulting from the two models, the term "RNA-STR" is used specifically for trees being constructed using the RNA7A model.

In the RNA trees, most classes lacked support but still remained in one clade, and only Classes A, B and D sometimes showed a support (Figure 29B-E). Two classes, F and G, more often appeared polyphyletic (Figure 29C, D, E), resembling that Class F was also polyphyletic as seen in trees resulted from the 2009 study (Figure 29A) [181]. Since Class F belonged to one clade using the global lenient RNA mask (Figure 29B, D), the

number of characters could have affected the tree topology. However, this contradicted the result for Class G, which appeared polyphyletic while using the larger global lenient mask under the RNA7A model (Figure 29D). Although with the same lenient mask but under the GTR model Class G appeared monophyletic, the model of choice may not have a major effect on the tree topology as no effect was seen for Class F between different models. Both Classes F and G are less robust than other classes in the RNA-based phylogeny, but should still be solid classes according to the taxon sampling tests (later section).

Unlike the global IEP-based trees, more disagreements were observed among the four global RNA trees. The most obvious disagreement was that Class A was internal to ML three times (Figure 29B, C, E). However, because there was still one tree that did not show such an arrangement, and Class A was always independent from ML in all taxon sampled trees (see "taxon sampling" section), this placement of Class A being internal to ML was suspicious and unconvincing. The second involved Class D, which was placed in multiple locations within the cluster of Classes A, B, G, ML and CL (Figure 29B-E). Similarly, Class D showed different locations in both taxon sampled trees and morphological trees (see later sections). If based on the overall trend observed in multiple trees constructed in this study, Class D was most frequently placed between two clusters, C+E+F and A+B+G+ML+CL, but never with firm statistical support. Regardless of more inconsistency was observed in the RNA trees, all classes generally form three large groups: C+F+E+unclassified, D+A+ML+G+Hp.au.I3 and B+CL1+CL2 (Figure 29B-E).

In comparison the topology obtained in 2009 [181], Class F was most different as it was grouped with B and ML in the 2009 study (Figure 29A) but was grouped in all four trees with C and E this time (Figure 29B-E). Indeed, Class F has never been placed close to either B or ML in trees inferred with a larger dataset after the 2009 study, suggesting that the data size could be the reason causing this topological difference for Class F. Possibly for the same reason, the location of the cluster of CL1+CL2 was also different between the 2009 study and new trees constructed this time (Figure 29). In addition, the current arrangement of CL1+CL2 has been continuously observed in trees constructed between the 2009 study and this study.

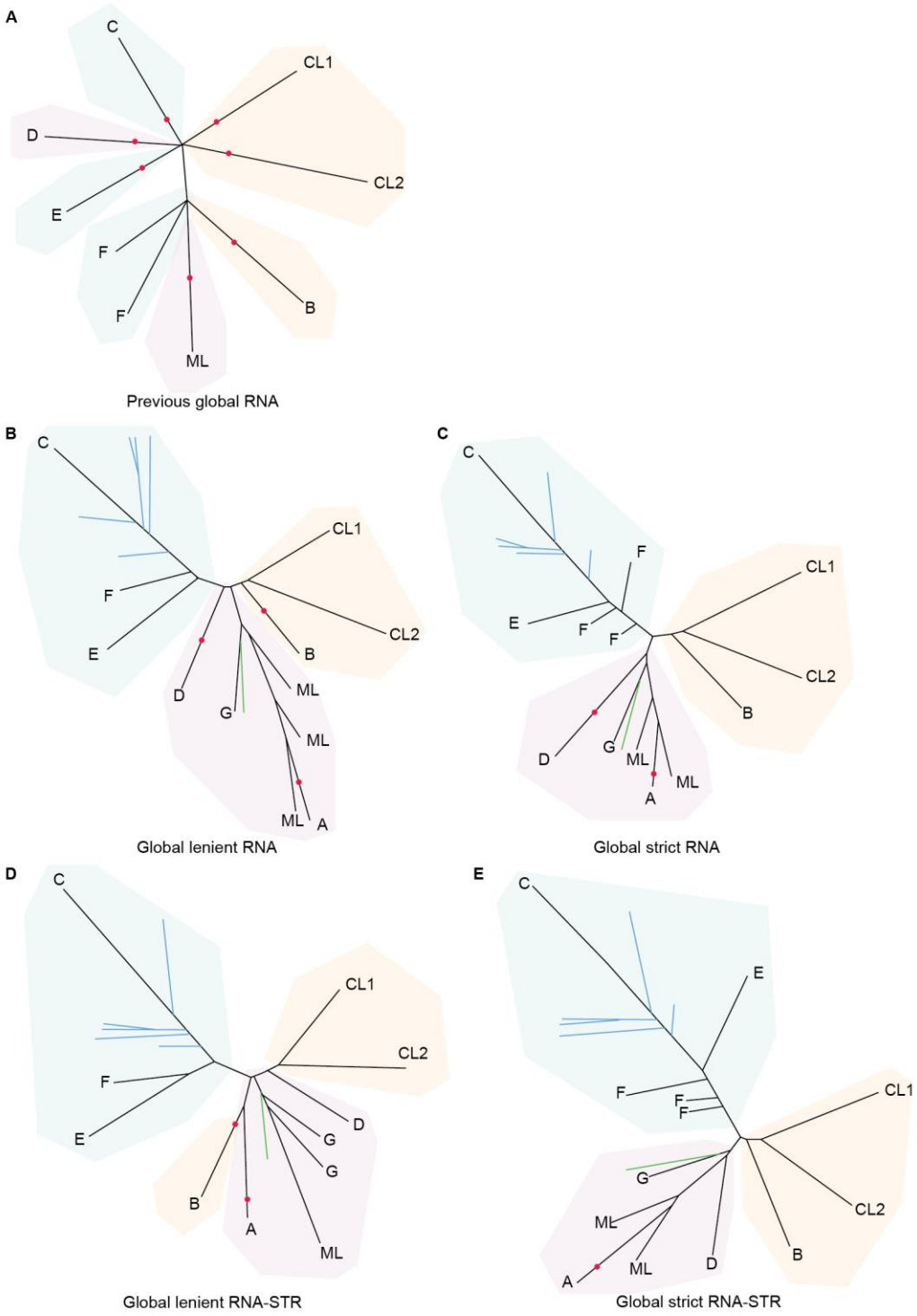


Figure 29. Global RNA trees inferred by various data types and masks.

Trees are simplified at the class or subclass level. Unclassified introns are drawn in blue except individual Hp.au.I3 that is in green. Nodes with a support (posterior probability >= 0.95 for panel A or bootstrap >= 75% for panel B-E) are indicated by a red dot. Coloured shadings correspond to the trend as in the main text. **A)** Adapted global RNA tree from the 2009 study [181]. **B and C)** Global lenient and strict RNA trees. **D and E)** Global lenient and strict RNA-STR trees. Tree in A) was constructed by MrBayes and trees in B-E) were constructed by RAxML.

Due to the small number of alignable characters (206 nt), it was anticipated that the RNA-based trees would suffer from low resolution, which was reflected by a ~20% decrease in resolution compared to the IEP-based trees (Figure 28). The RNA7A model seemed not to result in significant differences compared to the GTR model in terms of either the overall topology or the resulting resolution, even though trees constructed under the RNA7A model had lower resolution on average (Figure 28). Also because the GTR model is available in almost all programs designed for phylogenetic analyses, but the RNA7A model is only available in limited programs, the lenient RNA mask together with the GTR model were mainly used in following analyses.

Differences between IEP- and RNA-based topologies

As described in previous sections, the IEP- and RNA-based phylogenies both showed an overall similar trend in the relationships among classes, consistent with the hypothesis that group II intron IEP and RNA have coevolved. However, there were several disagreements for specific subsets of classes. First, in the IEP phylogeny, classes CL1 and CL2 each split into two supported subclades (CL1A, CL1B, CL2A, CL2B) while CL2 was internal to CL1 (Figure 27). The four subclades showed multiple arrangements, but CL1B was always the earliest branching clade compared to the others (Figure 27). In contrast, in the RNA phylogeny, although members belonging to the same subclade were usually closely related, no subclades could be identified (Figure 29). Moreover, CL1 and CL2 were independent of each other, appearing as sister clades (Figure 29). These differences were also observed in the 2009 study [181], indicating the enlarged dataset did not contribute to resolve the relationship between the two CL classes. Whether CL1 and CL2 have non-coevolving histories between their IEP and RNA remains unclear. Because CL1 and CL2 generally are more similar, this question was further investigated by constructing CL-specific trees (see next section).

Similar to the CL subclades, the C1 and C2 subclades from Class C could be observed in the IEP-based phylogeny but not in the RNA-based phylogeny (Figure 27, Figure 29). Although members of the early branching clade C1 were mostly found close to each other in the RNA-based trees, they did not form one single clade and they were not always the earliest branching taxa. However, this might be caused by the divergent sequences of Class C ribozymes as the data size increases, because the later taxon sampled trees showed supported C1 and C2 subclades using the RNA sequence with

smaller datasets (taxon sampling section).

The next disagreement involved the closest relative of the CL1+CL2 clade, which was Class G in the IEP-based tree but Class B in the RNA-based tree, even though without support (Figure 27, Figure 29, Figure 30). This situation correlated with the unclassified intron Hp.au.I3 and its closest relative. Hp.au.I3 was always within the cluster of A+B+ML in the IEP-based trees, but was always grouped with Class G in the RNA-based trees and sometimes even with support (Figure 27, Figure 29, Figure 30). Moreover, it was indeed noticed that the RNA sequence of Hp.au.I3 can be aligned well with other Class G introns, although its IEP sequence always appeared to be unclassified. Because almost all Class B introns are from Firmicutes while almost all Class G introns are from Actinobacteria and Proteobacteria (no Class G introns are from Firmicutes and no Class B introns are from Actinobacteria and Proteobacteria either), it would be less likely for these two classes to undergo convergent evolution judging from they have different cellular environments. Therefore, it would be more likely that there was a sequence swapping event of the IEP occurred between ancestors of these two classes, which resulted the two classes having switched locations in the IEP- and RNA-based trees respectively (Figure 30).

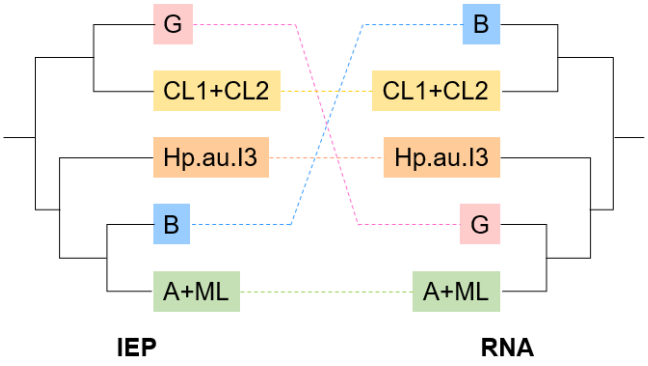


Figure 30. Simplified relationship between classes G and G and their neighbour classes.
Topologies shown in this figure were derived from global IEP- and RNA-based trees using the lenient masks. Trees are only intended to illustrate the relationship between classes and are not drawn to scale. If both trees reflect the actual evolutionary history, it was likely that an IEP swapping event was occurred between ancestors of Classes B and G, resulting their switched locations in the IEP- and RNA-based trees.

The last difference concerned Class D, which often remained between the two clusters of classes C+F+E and A+B+G+ML+CL. However, it was more often observed that Class D was grouped with the former cluster in the IEP-based trees, and with the latter cluster in the RNA-based trees (Figure 27, Figure 29). As sometimes a support could be observed (in taxon sampled trees), this may indicate the RNA and IEP of class D introns generally were coevolving, but could have a different evolutionary rate that resulted in different groupings when these are based on different types of data.

Class-specific trees

As the global trees included all 661 introns from 10 different classes, many areas of both the IEP and RNA alignments that do not have sequence or structural similarity had to be excluded, even though they may be conserved within specific classes. Therefore, class-specific trees were expected to be able to provide a higher resolution, as class-specific sequence alignments should include more alignable characters. In this project, class-specific trees were constructed for all 10 classes using both the IEP and RNA data, and also multiple class-specific sequence masks (Table 4B).

Class-specific masks included 100-200 more aa than global masks (Table 4B). Among masks for the same class, a higher resolution was usually obtained while using a larger mask (Figure 31, Figure 32), which was consistent with the trend seen from global masks (Figure 28). All trees at the class level, including both individually constructed class-specific trees and subtrees extracted from global trees, were then subjected to pairwise tree comparisons performed manually (see later "topology test" section), which only compared supported clades and showed no significant differences between the IEP- and RNA-based trees. Therefore, these class-specific trees were concluded to be more accurate in representing the relationship between individuals of each class, while the class-specific lenient IEP and RNA trees were determined to be the best tree within each class.

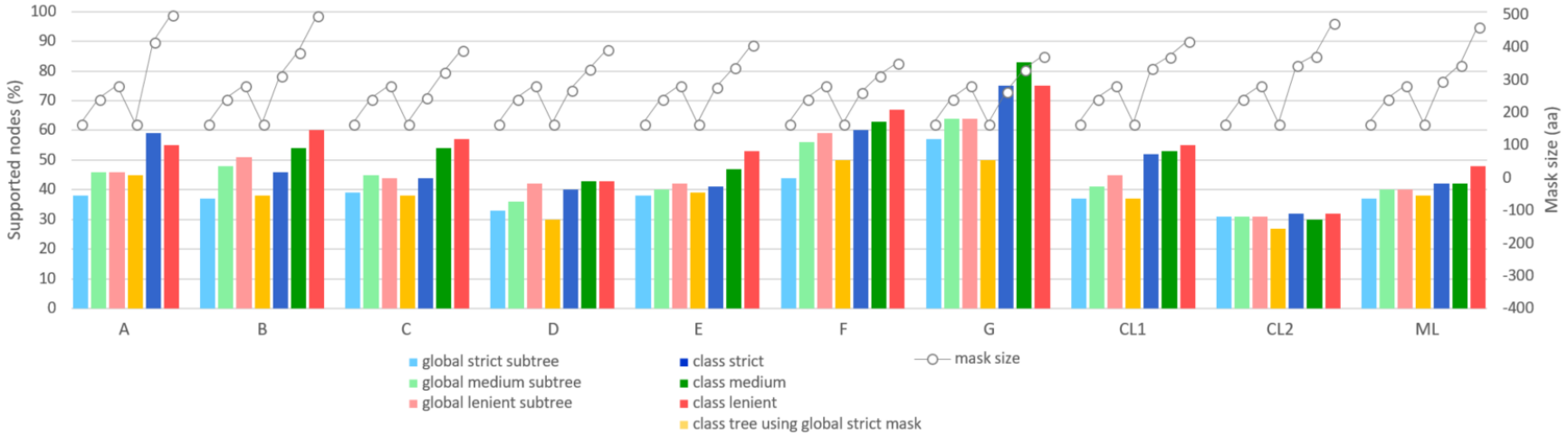


Figure 31. Percentage of supported nodes and mask size of IEP-based trees at the class level.

The first y-axis (left) corresponds to the percentage of supported nodes (bootstrap >= 75%) for each IEP-based tree. Trees are either extracted from global trees or individual class-specific trees. Note that "global strict, class-specific" trees were constructed using the global strict IEP mask. The second y-axis (right) corresponds to the mask size in a unit of aa.

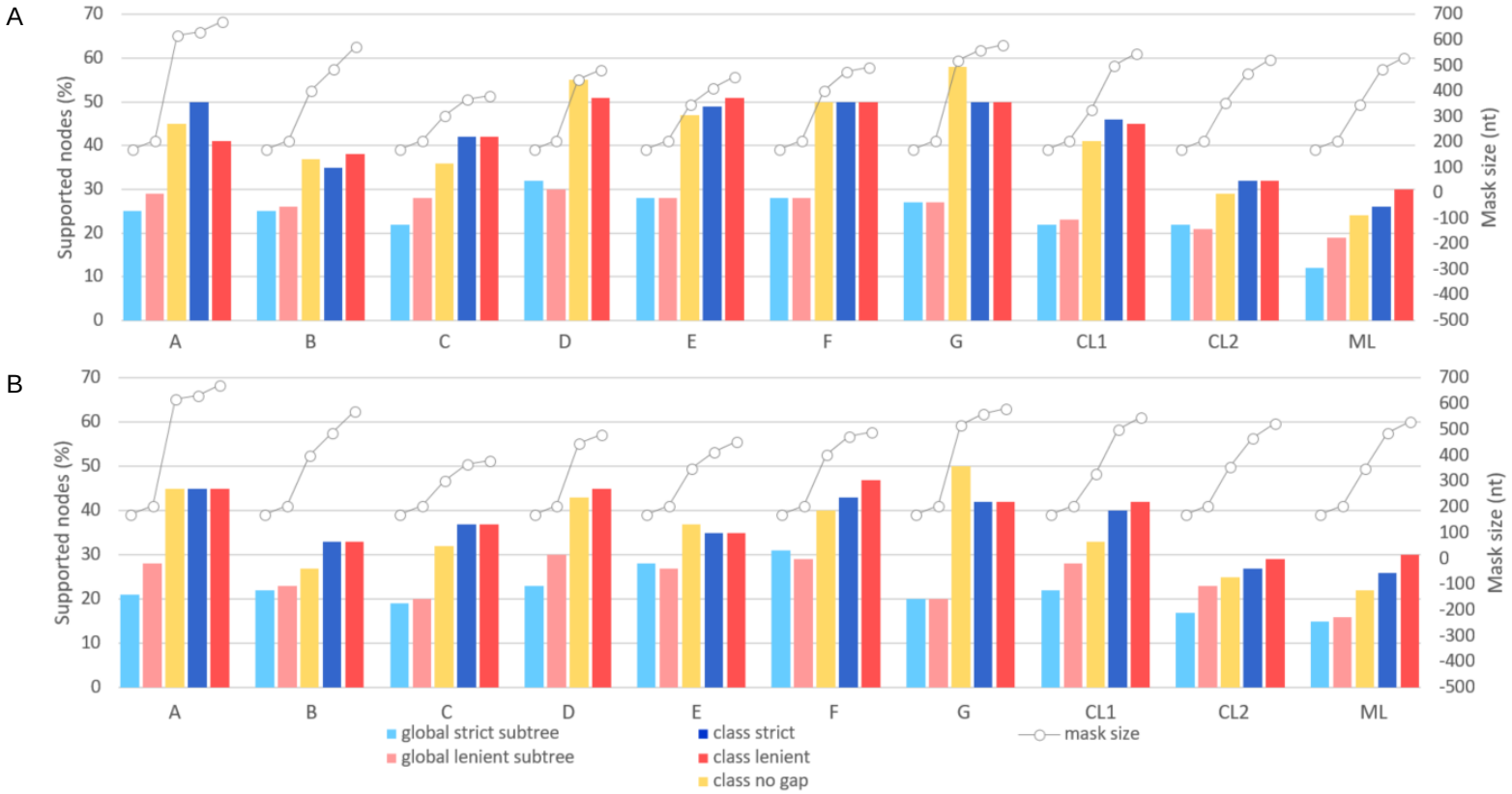


Figure 32. Percentage of supported nodes and mask size of RNA- and RNA-STR-based trees at the class level.

The first y-axis (left) corresponds to the percentage of supported nodes (bootstrap >= 75%) for each IEP-based tree. Trees are either extracted from global trees or individual class-specific trees. Note that "global strict, class-tree" was a class-specific tree constructed using the global strict IEP mask. The second y-axis (right) corresponds to the mask size in a unit of aa. Note that Class D does not have a class-specific strict RNA mask because RNA sequences of this class are highly alignable. **A)** Trees of the RNA set. **B)** Trees of the RNA-STR set.

Next, to compare whether the tree topology could be affected by using the same mask but different sets of taxa, the global strict IEP mask (168 aa) was used as a test mask. For each class, a class-specific tree was constructed using this mask and then compared with the global IEP tree using the same mask. Overall, both trees showed similar resolutions for all classes (Figure 31, light blue and yellow bars), and their topologies did not reveal any disagreements either (see later "topology test" section), which indicated that the addition of extra taxa (e.g. introns from other classes) would not affect the tree when the same mask is used.

Since class-specific trees have been concluded to be the most accurate, the issue involving a disagreement in topology between CL1 and CL2 as mentioned in the above section could be investigated by constructing CL-specific trees. IEP and RNA alignments including only CL1 and CL2 introns were created and CL-specific masks were created (not shown), which had an increase of ~110 aa and ~100 nt compared to the global lenient masks. The resulting CL-specific tree using the IEP data was very similar to the global tree. The only difference was that CL1B was not monophyletic (figure not shown).

In contrast, the CL-specific RNA tree was different from the global lenient RNA tree but agreed with the IEP-based topology (figure not shown). Instead of being sister clades as seen in the global RNA tree, CL2 was internal to CL1 in the CL-specific RNA tree, although without a support (figure not shown). However, none of the four subclades could be identified in the CL-specific RNA tree, although members from the same subclade tended to be closely related, and all members from CL1B were branching the earliest, consistent to the observation as of the global IEP tree.

In conclusion, the CL-specific trees agreed more with each other as well as with the global IEP-based topology, indicating that the IEP and RNA of the two CL classes coevolved. Therefore, disagreements observed between global trees were likely due to the limited number of characters of global masks. However, the approach of using class-specific trees could not be applied to resolve other disagreements observed in the previous section, because other classes (B, D and G) do not share more characters compared while being aligned together to the global masks.

Factors that may affect tree topologies

Taxon sampling

Since the 2009 study reported that the trees are sensitive to taxon sampling, in which similar sets of taxa may result in different topologies, this was also investigated in this study. Four subsets of introns were evenly sampled along the global lenient IEP tree from each supported clades. Each contained approximately 140 taxa, which included ~20% taxa from each class. All six unclassified introns were included in all four subsets (Appendix D1).

Taxon sampled trees were inferred by both RAxML and MrBayes using the global lenient masks. Although both programs generated nearly identical topologies, trees inferred by MrBayes showed a larger number of supported nodes. Using the IEP data, all four subsets showed consistent topologies among themselves and only had minor disagreements involving the branching pattern of ML, and the relationship between subclades CL1A, CL2A and CL2B (Figure 33A). The overall topology of these sampled IEP trees was consistent with the global IEP trees. Supports could be observed for all classes, including Classes C and F that were unsupported in global IEP trees (Figure 33A), indicating that Classes C and F are both monophyletic and solid classes, even they lacked support in global IEP trees.

More surprisingly, IEP-based trees inferred by MrBayes even had supports beyond many classes, including D+E, G+CL, A+B+ML and some larger clusters (Figure 33A). Even though the 2009 study also used the Bayesian method to infer trees, supports beyond classes were never observed. Because taxon sampled trees inferred this time used the same IEP mask as used in 2009, the use of a large dataset was thus believed to be the main reason of obtaining resolutions beyond classes.

In addition, the 2009 study had a lot of repeating taxa in different subsets due to the small dataset [181], which might cause an artificial bias. But this time, all four subsets had mostly unique taxa (Appendix D1), and thus should equally represent the entire population. Altogether, as no major differences were observed across the four subsets, the sensitivity to taxon sampling could thus be eliminated by using a larger dataset, and these trees also provided more evidence of the relationship beyond classes.

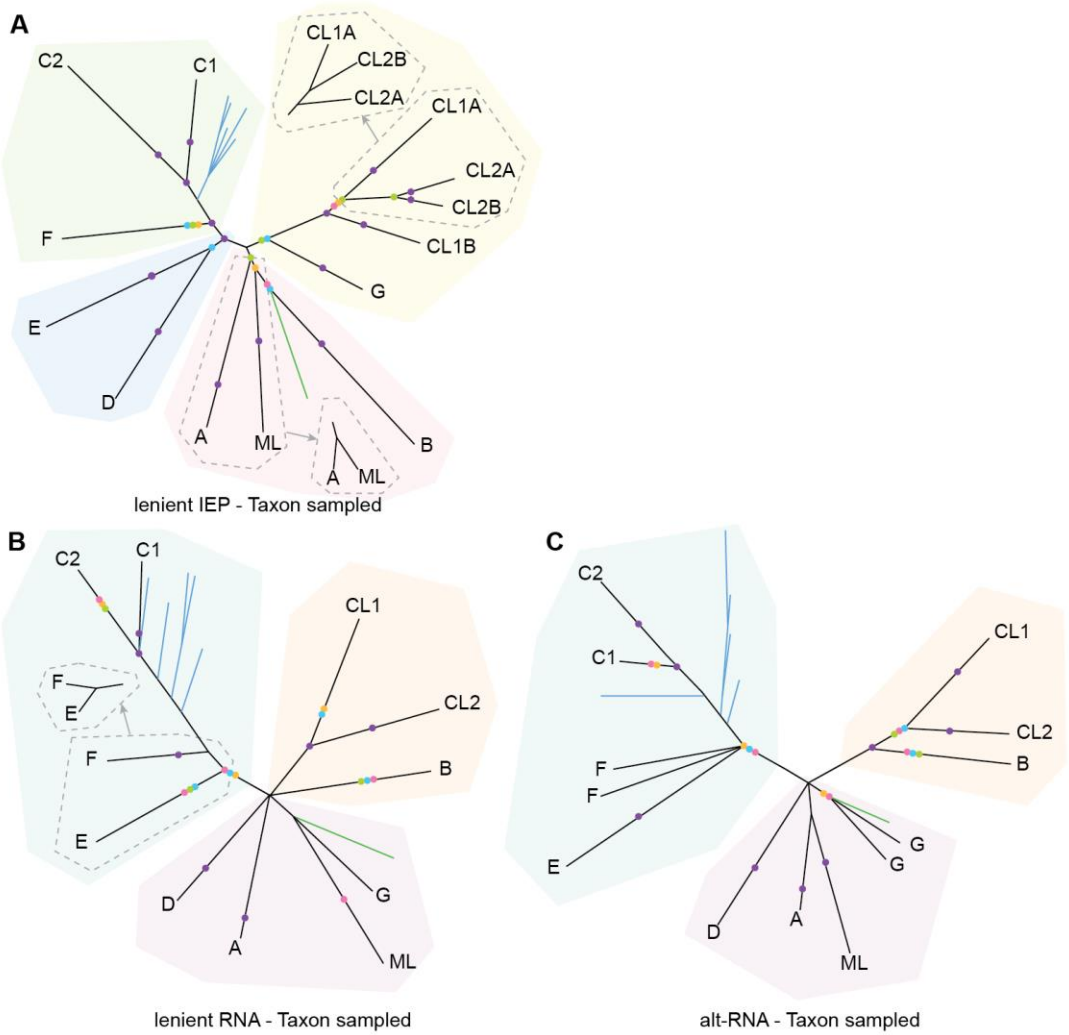


Figure 33. Consensus trees of four taxon sampled subsets.

Four subsets of group II introns were used to test the sensitivity to taxon sampling. Each subset consisted ~20% of introns from each class plus all six unclassified individuals. Trees were constructed by both RAxML and MrBayes. All four subsets showed similar topologies and were summarised onto one consensus tree using: **A)** the global lenient IEP mask, **B)** the global lenient RNA mask, and **C)** the alt-RNA mask. Supported clades (posterior probability >= 0.95) are indicated by coloured dots. while pink, orange, blue and green correspond to subsets 1-4 respectively, purple indicates a node was supported in all four trees. Because Bayesian trees had more supports including those between classes, this figure only plots supported nodes observed from Bayesian trees. Unclassified individuals are drawn in blue except Hp.au.I3 is in green. Alternative common arrangements are circled by dashed grey lines. Shadings in different colours correspond to the trend as described in the main text.

In the 2009 study, the effect of taxon sampling was only tested using the IEP sequences. Because the larger dataset seemed not to be sensitive to taxon sampling using the IEP data, the RNA-based taxon sampling tests were also performed. In addition to the global lenient RNA mask, another mask, named "alt-RNA", which covered the same areas as

used previously in 2009 (Figure 25) [181] was also included in order to compare whether the different RNA masks used in the two projects could affect the tree topology. The alt-RNA mask covered 138 nt (Table 4) and was created mainly based on the sequence similarity, while the global lenient RNA mask covered 206 nt (Table 4) and considered homologies in both sequence and secondary structure. Some crucial interactions, such as the α-α' and EBS1-IBS1 interactions (Figure 25A) were not included previously due to the lack of sequence similarity, but were included this time as structurally conserved motifs. In contrast, the ε' motif (Figure 25A) was included in the 2009 study but not this time because their secondary structures vary among classes.

Trees using these two RNA masks did not reveal significant disagreements (Figure 33B, C). In general, these sampled RNA trees showed more supports at the class level, as was seen from the sampled IEP trees. However, similar to global RNA trees, Classes F and G were still unstable compared to other classes and sometimes were polyphyletic (Figure 33B, C). Unlike sampled IEP trees, however, supports beyond classes were rarely observed, and the only support was observed to separate the two clusters of C+E+F+unclassified and A+B+D+G+ML+CL (Figure 33B, C). This agreed with the trend seen in global RNA trees by grouping class D together with A, B, D, G, ML and CL, although the relationship between these classes were often unclear. However, Class A was never seen internal to ML (Figure 33B, C), implying the earlier observation in the global RNA trees (Figure 29B, C, E) was possibly untrue. On the other hand, the cluster of B+CL was supported in all four trees using the alt-RNA mask (Figure 33C), while the same arrangement was observed in the global RNA trees but was supported (Figure 29B, C, E).

Different from all global RNA trees, Class C split into subclades C1 and C2 with a support, while C1 was still the earliest branching clade (Figure 33B, C). Class C as a whole was supported in all four trees, indicating the disagreement observed in global RNA trees involving Class C could be due to the poor resolution, and both the IEP and RNA of Class C were concluded to be coevolving.

In conclusion, with the larger dataset, trees were not sensitive to taxon sampling for either the IEP or RNA datasets. Taxon sampled trees generally agreed with global trees, and some supports between classes were observed through Bayesian inference especially for the IEP dataset. It could be expected that Bayesian trees involving all 661 taxa may still yield supports beyond classes, but to construct such trees seemed less

feasible as it would require huge amount of computational time and resource (of note, preliminary tests have shown to estimate a Bayesian tree for Class C only would take more than two months). Although the two different RNA masks did not show conflicting topologies, the resolution of RNA-based trees was not improved as much as IEP-based trees, and the lack of alignable characters was the main issue for the RNA dataset.

GC content biased sampling

Base composition heterogeneity was another factor that could affect the tree topology as suggested in the 2009 study, in which unrelated taxa could be brought together during phylogenetic inference if they share a similar base composition. Base composition heterogeneity is often reflected by the GC content. A wide range of GC content for the IEP, from 26 to 68%, was observed across all 661 introns used in this project (Figure 34A). Therefore, to investigate whether the tree is sensitive to base composition heterogeneity with the larger dataset, a total of 196 introns with a similar GC content (43-52% inclusive, Figure 34A) was selected as an "GC unbiased" subset (Appendix D1), and both IEP- and RNA-based trees were constructed. Since Class G generally have higher GC content, only one intron (Al.fi.I1) was included in this subset. To test whether Class G was still monophyletic, four additional Class G introns were appended into the GC unbiased subset. One of the four has a 58% GC while the others have 62% GC. However, this did not affect the location of Class G in both IEP- and RNA-based trees, and the five Class G introns were always monophyletic with a support (data not shown).

The IEP-based tree using the GC unbiased subset showed most classes being consistent with the result from the global IEP tree, but the most obvious difference was Class A being placed between C and F instead of its most common position near ML (Figure 35A). When being tested using the other two masks (medium, strict), the same arrangement was still observed (data not shown), which confirmed the location of Class A was likely affected due to the GC unbiased dataset. In contrast, the RNA-based trees were generally consistent to the global RNA trees, although there were still a few differences (Figure 35B, C). While using the global lenient RNA mask, only Class B was located differently as it was not grouped with CL1 and CL2, although still close to them (Figure 35B); when using the alt-RNA mask, three differences were observed, including that Class F was internal to E, Class B was placed between CL1 and CL2, and Class D

was internal to the cluster of A+G+ML (Figure 35C).

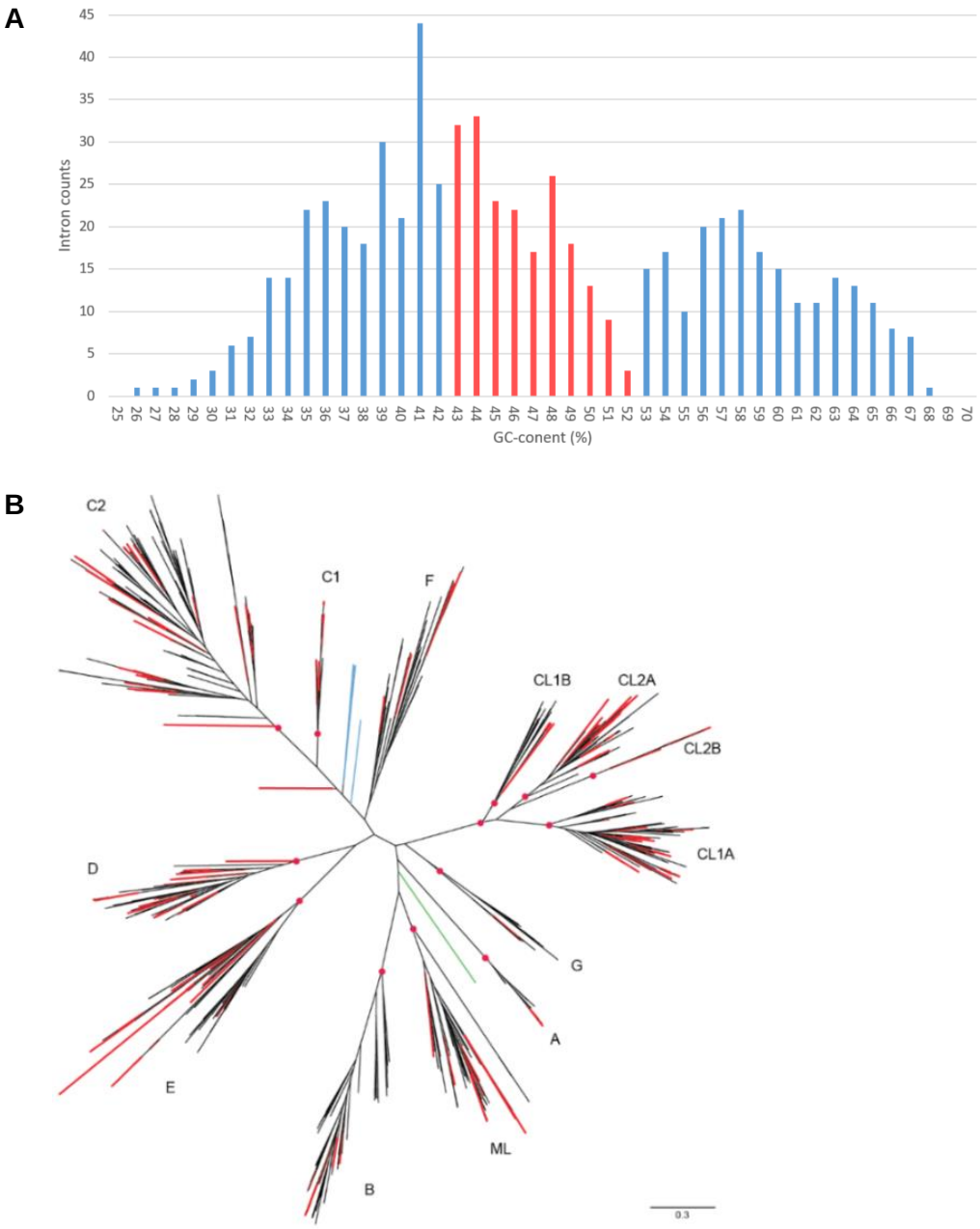


Figure 34. The GC-content based subset.

A) Counts of introns of various GC contents of the IEP. Red bars correspond to the range included in the GC-content based subset. **B)** The distribution of introns included in the GC-content based subset (red), which also corresponds to red bars in A), mapped onto the global lenient IEP tree. Classes with a >= 75% bootstrap support are indicated by a red dot. Blue branches correspond to unclassified introns except Hp.au.I3 which is in green.

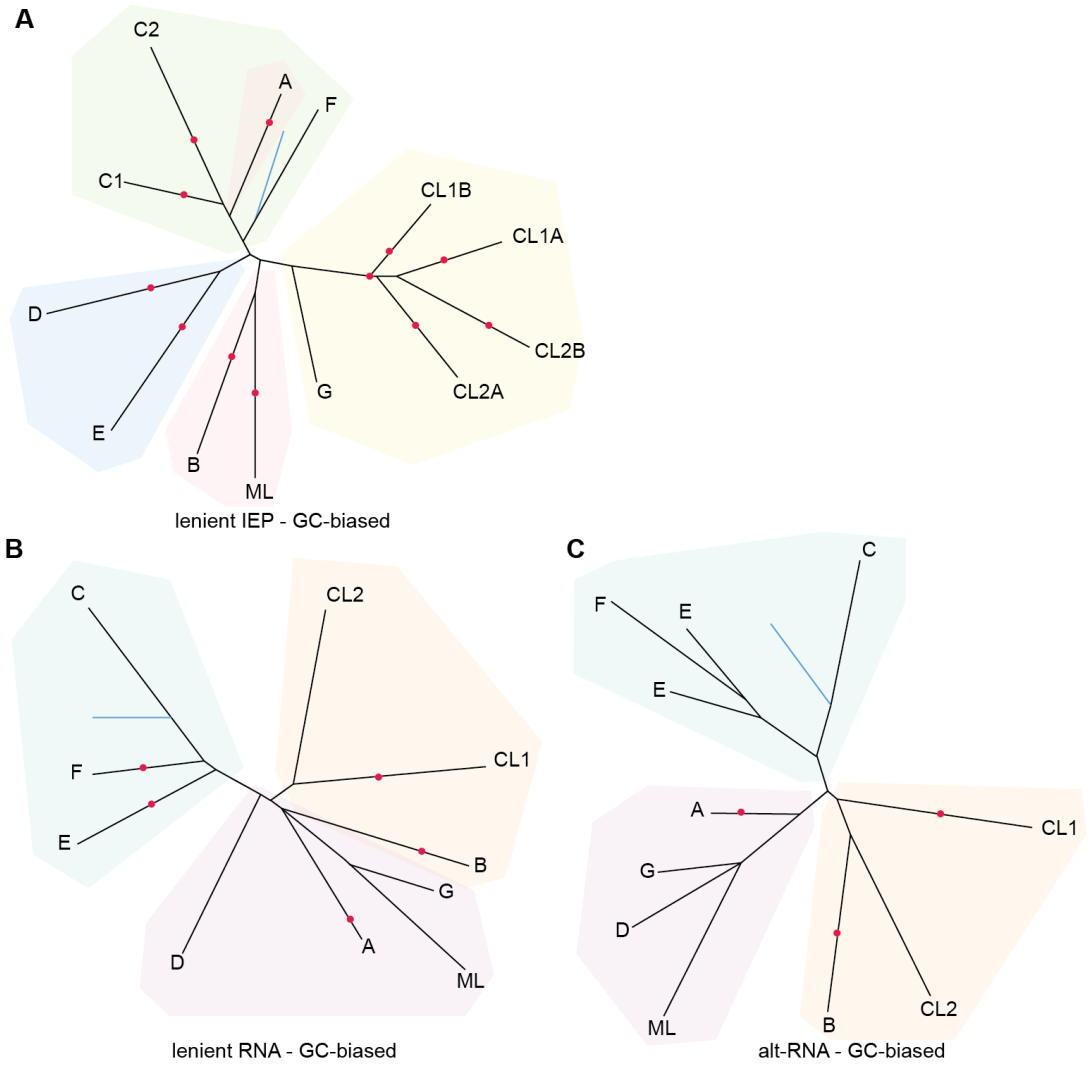


Figure 35. Trees constructed using taxa with similar GC contents.

The GC-unbiased subset consisted 196 taxa with a GC content ranging from 43-52% and were constructed by RAxML using **A)** the global IEP lenient mask, **B)** the global RNA lenient mask, and **C)** the global alt-RNA mask. A red dot indicates a high bootstrap support (>= 75%). Shadings in different colours correspond to the trend as in main text. Note that only one Class G intron was involved and thus Class G was unable to be evaluated for the class-wide confidence. Blue line corresponds to intron UA.I4, which was the only unclassified intron within the selected range of GC content.

Even though only Class A showed a difference while using the IEP data, it still seems that the GC content does affect the tree topology. To provide additional evidence, other GC unbiased samplings were created by including introns from different ranges of GC content (e.g. 10-20%, 20-30%, etc.). As a result, almost all these GC unbiased samplings showed unexpected topologies for both the IEP and RNA data, while between different subsets no consistent topologies could be observed either (data not shown). It

was clear that the base composition heterogeneity affects the tree topology, and there could be bias due to the GC content when the dataset is small and does not represent the whole population. Therefore, the more samples being included in the dataset, the closer the dataset could reflect the actual situation of all introns, and thus to diminish the sensitivity to the base composition heterogeneity.

Topology tests to compare IEP and RNA evolution

Observations so far implied the IEP and RNA were likely coevolving. In order to obtain a more formal evaluation, the Shimodaira-Hasegawa (SH) topology test was performed. In the 2009 study, the SH test between IEP- and RNA-based trees indicated there was likely a conflict between the IEP and RNA, even though it could not be identified [181]. This time, SH tests equivalent to that of the 2009 study were first performed to compare global trees, including all four sets of the taxon sampled trees constructed using the lenient IEP, the lenient RNA and the alt-RNA masks as described earlier. It showed the same results as the 2009 study, that the IEP- and RNA-based trees were conflicting and likely not coevolving (data not shown).

Next, SH tests were used to compare class-specific trees, which had higher resolutions and were believed to be more accurate. All trees at the class level, including the class-specific trees using different masks and subtrees extracted from the global trees using different masks, were compared. Between trees constructed from the same data type (IEP vs IEP, RNA vs RNA), the SH test usually passed (Table 5, Appendix D3). However, there were still exceptions, in which the strict IEP trees were rejected by the SH test when they were compared with other IEP trees. This was observed for most classes, including A, B, C, E, CL1, CL2 and ML. For all classes, the two global RNA trees (lenient and strict) were rejected by the SH test when they were compared with other RNA trees (Table 5, Appendix D3). While comparing trees constructed from different data types (IEP vs RNA), most SH tests rejected the null hypothesis, and usually resulted in a p-value being either 0 or close to 0 (Appendix D3). The only exceptions were Classes A and G, which showed more than half of the trees made from different data types passed the SH test (Table 5, Appendix D3).

Table 5. Results of SH tests for Class A.

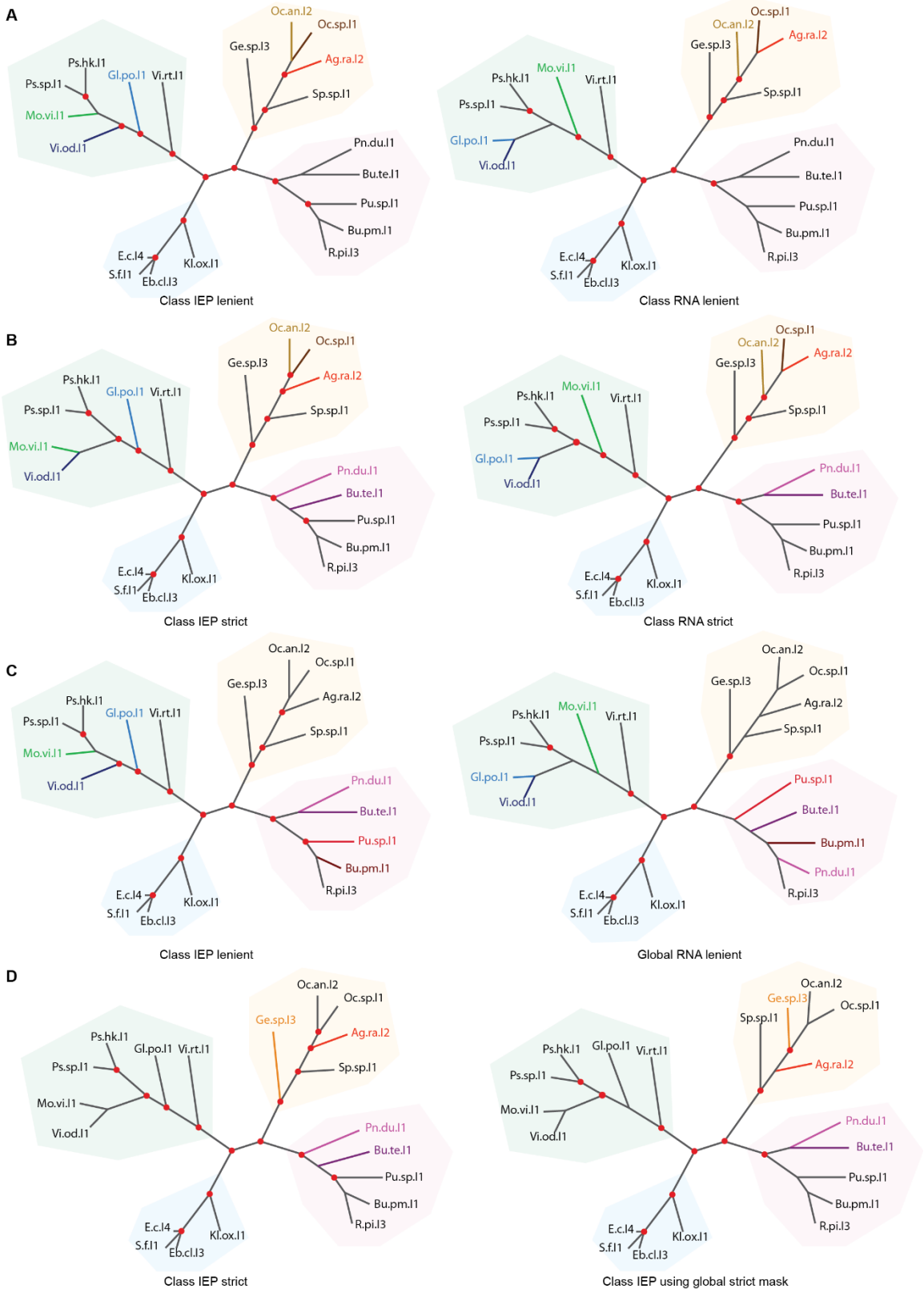
The SH test was performed by programs PAML and CONSEL between the same list of trees and various masks. The significance level was set to α=0.05. A plus sign "+" indicates values higher than the significance level (fail to reject the null hypothesis). Red shadings indicate the rejection occurred between trees and alignments of the same data type (IEP vs IEP, RNA vs RNA), and blue shadings indicate the rejection occurred between different data types (IEP vs RNA). Results of other classes are provided in Appendix D3. **A)** Before removing the suspected conflicting taxon Ge.sp.I3. **B)** After removing Ge.sp.I3. Note that global trees of 661 taxa were not reconstructed after removing Ge.sp.I3, and were thus not included in the second round of SH tests.

	IEP alignment			RNA alignment		
Tree list	**class lenient**	**class strict**	**global strict, class tree**	**class lenient**	**class strict**	**class no gap**
A. Before removing Ge.sp.I3						
IEP class lenient	0.955+	0.925+	0.608+	0.181+	0.156+	0.127+
IEP class strict	0.896+	0.938+	0.567+	0.124+	0.094+	0.109+
IEP global lenient	0.518+	0.430+	0.450+	0.029	0.025	0.021
IEP global medium	0.507+	0.525+	0.701+	0.058+	0.043	0.055+
IEP global strict	0.522+	0.523+	0.928+	0.029	0.027	0.028
IEP global strict, class tree	0.017	0.013	0.905+	0.02	0.018	0.02
RNA class lenient	0.185+	0.260+	0.045	0.996+	0.998+	0.992+
RNA class no gap	0.185+	0.260+	0.045	0.995+	0.969+	0.996+
RNA class strict	0.074+	0.117+	0.045	0.992+	0.998+	0.994+
RNA global lenient	0.002	0.002	0.009	0.155+	0.097+	0.090+
RNA global strict	0.0002	0.0002	0.004	0.003	0.003	0.004
RNA-STR class lenient	0.065+	0.105+	0.045	0.988+	0.993+	0.989+
RNA-STR class no gap	0.160+	0.229+	0.043	0.792+	0.789+	0.816+
RNA-STR class strict	0.055+	0.084+	0.043	0.792+	0.790+	0.816+
RNA-STR global lenient	0.003	0.003	0.006	0.253+	0.193+	0.181+
RNA-STR global strict	0.005	0.004	0.012	0.488+	0.369+	0.324+
B. After removing Ge.sp.I3						
IEP class lenient	0.924+	0.919+	0.665+	0.057+	0.049	0.042
IEP class strict	0.786+	0.804+	0.299+	0.056+	0.047	0.041
IEP global strict, class tree	0.488+	0.544+	0.939+	0.047	0.045	0.041
RNA class lenient	0.025	0.055+	0.035	0.985+	0.973+	0.974+
RNA class no gap	0.025	0.055+	0.035	0.968+	0.976+	0.987+
RNA class strict	0.025	0.055+	0.035	0.988+	0.968+	0.958+
RNA-STR class lenient	0.149+	0.189+	0.150+	0.516+	0.470+	0.485+
RNA-STR class no gap	0.019	0.04	0.033	0.622+	0.623+	0.671+
RNA-STR class strict	0.019	0.04	0.033	0.622+	0.623+	0.671+

The results of SH tests obtained so far, were intriguing to interpret. It could be reasonable that the IEP- and RNA-based trees were rejected by the SH test because they were not coevolving. Trees constructed using the same type of data (e.g. all IEP-based trees) would be expected to have similar topologies and should always pass the SH test, because they were constructed using the same sequence alignment, even though with various masks. It was generally true that most trees made from the same data type passed the SH test, but there were exceptions as stated above.

Therefore, Class A was used as an example to investigate conflicts indicated by the SH test. This was not only because Class A is a relatively small class containing 25 taxa in the dataset, but also because both its IEP and RNA sequences are conserved and can be aligned without ambiguity. The Class A specific lenient masks for IEP and RNA had 502 aa and 673 nt in length, which were 217 aa and 467 nt longer than the respective global lenient masks (Table 4).

Between various IEP and RNA trees for Class A, the best trees (class lenient IEP and class lenient RNA) passed the SH test (Table 5A). A visual topologic comparison that examined the locations of all taxa, the two trees were indeed consistent in the tree topology (Figure 36A). But other IEP and RNA trees did not always pass the test, even though the tree topologies did not appear inconsistent (Figure 36B, C). Interestingly, if strictly relying on the bootstrap support, some trees that passed the SH test could even contain conflicting nodes, in which each was supported but contained different sets of taxa (Figure 36A, B, green sector). In contrast, some trees that were rejected by the SH test contained no conflicting nodes at all (Figure 36C, green and pink sectors). Finally, through the manual tree comparison, it could be seen that even when two IEP trees were rejected by the SH test, the observed topological differences could still be considered minor as only a small local area of the entire tree was involved (Figure 36D, yellow sector).



(Figure caption on the next page.)

Figure 36. Selected examples of tree comparisons for Class A.

Each panel compares two trees, and the name is given below each tree. Red dots indicate supported nodes with a bootstrap value >= 75%. Four coloured sectors (green, yellow, pink, blue) correspond to four major clades within Class A. Coloured branches and intron names corresponded to topologic differences between the two trees while the node support was not taken into account. Note that the node containing E.c.I4, S.f.I1 and Eb.cl.I3 in the blue sector is simplified in this figure. It actually contains eight taxa in total, and all taxa share identical sequences for both IEP and RNA. Trees are not drawn to scale. **A, B)** The SH test failed to reject the null hypothesis; **C, D)** The SH test rejected the null hypothesis.

As illustrated in Figure 36D, the manual comparison showed two introns, Ge.sp.I3 and Pn.du.I1, have different arrangements between the strict class A IEP tree and the class A IEP tree using the global strict IEP mask, and both could potentially cause the conflict between the two trees. However, because Pn.du.I1 in both trees only differed by the bifurcating pattern with its closest relative Bu.te.I1, this was considered as minor. In contrast, Ge.sp.I3 was placed at different locations that also affected its closest relatives, this intron was considered to be more likely to cause the conflict. Therefore, Ge.sp.I3 was removed and new sets of Class A trees were constructed. This time, all trees constructed using the same data type passed the SH test, but over half the comparisons between IEP and RNA trees were still rejected (Table 5B). Nonetheless, manual tree comparisons across these new trees did not reveal significant differences between conflicting IEP and RNA trees, only minor disagreements resembling examples in Figure 36.

Among rejected trees constructed from the same data type, the strict mask was always involved (Appendix D3), implying strict constraints in sequence alignment, which retained only the most highly conserved characters, could result in a different topology that may not reflect the actual evolution, and could cause the SH test to reject the null hypothesis. Group II introns are known to have a large sequence divergence especially of the RNA ribozyme. As seen in the RNA-based trees, the same set of taxa may show various arrangements in different trees but none was supported. Therefore, it may not be suitable to evaluate such trees using standard topology tests, which are likely to be rejected due to unsupported and minor disagreements.

To investigate whether this situation, in which only small clades being affected while the overall topology remained consistent, was common in other classes, manual comparisons were done across all trees for all classes. However, the result was similar to what was observed for Class A and only included minor disagreements that were less

possible to be considered as conflicts (data of minor disagreements not shown). Of all classes, only two sets of introns from Class C were suspected to be potential conflicts, because they were more often observed and affected a rather large area in the Class C trees (Figure 37). Nonetheless, the affected area only takes up a small part out of the entire Class C tree, and thus disagreements caused by these introns may be still considered as minor. In conclusion, the manual tree comparisons indicated that the IEP and RNA trees were consistent, and the IEP and RNA of most (if not all) introns coevolved.

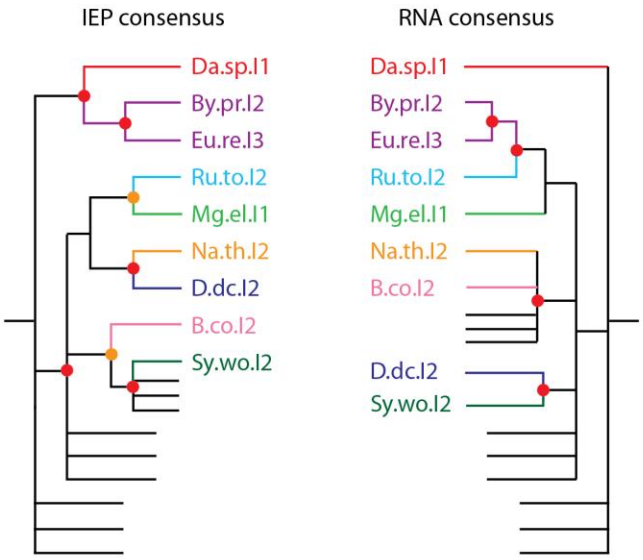


Figure 37. Potential topological conflicts between IEP and RNA trees for Class C.
Simplified consensus trees are drawn based on all trees constructed using the same data type (IEP and RNA), and only include a small local area where potential conflicts were observed. Red or orange dots are placed at nodes with a >= 75% bootstrap support, and red dots indicate a node is always supported in all trees of the same type, orange dots indicate the node is only supported in some trees of the same type. Affected introns are colour coded. Unlabelled branches in black represent unaffected introns, and the number of branches does not correspond to that of introns actually being omitted. Trees are not drawn to scale.

Morphology based topology

All trees inferred so far in this project only used models for nucleotide or amino acid sequences. However, large portions of both the IEP and RNA alignments were excluded from analysis due to the lack of sequence or structure similarity. This especially affected the character size used for the RNA-based phylogeny. A normal group II intron RNA

(excluding DIV and the IEP) is 800-1000 bp in length on average, but even the longest lenient global mask could only cover 206 nt (Table 4), which was merely a quarter of a normal full-length intron RNA. The lack of characters was believed to be the main reason causing a low resolution for all RNA-based trees. Regardless, these dismissed areas still have morphologic information, such as the length, the presence of a secondary structure or tertiary interaction, which can be used in phylogenetic analysis.

To recover such dismissed "morphological" information, both the IEP and RNA were examined while treating each class as one taxon and 135 morphological characters were collected. Thirty-seven (37) characters were from the IEP data based on the length and pairwise HMMER comparison, and 98 characters were from the RNA data based on structural features (Methods, Appendix D2). Using binary and morphological models available in MrBayes, four morphologic trees were constructed using characters from either or both IEP and RNA sources, or a subset being considered of "more importance" (Figure 38, Appendix D2, Methods).

Of the four morphologic trees, three showed similar topologies agreeing with the trend observed from global trees (Figure 38A, C, D), and the exception was the IEP-only tree. Of the three trees, supports were sometimes observed for clusters of C+B.ce.I2 (unclassified intron), C+unclassified, E+F+C+unclassified, and that D was separated from other classes (Figure 38A, C, D). The clade of CL1+CL2 was always supported even in the IEP-only tree, but the relationship between CL1 and CL2, which differed between the global IEP- and RNA-based phylogenies, could not be resolved because each class was treated here as one single taxon. The relationships between CL1+CL2 and B or G also remained unclear. In contrast, Class D was grouped with C+E+F while using all 135 characters but was separated from them if only using the RNA data, agreeing with the trend observed from global trees.

In contrast, the IEP-only tree was very different from the tree from the global IEP tree (Figure 38B). This was not surprising as the set of 37 characters was small and may not be enough to infer an accurate evolutionary pathway. Although the set of 49 "more important" characters was not of a high number either, they were specifically selected characters that corresponding to essential interactions or motifs, and thus were more informative. Thus, the IEP-only tree (Figure 38B) was not further considered, while other morphological trees can provide additional support to the unsupported relationship between classes.

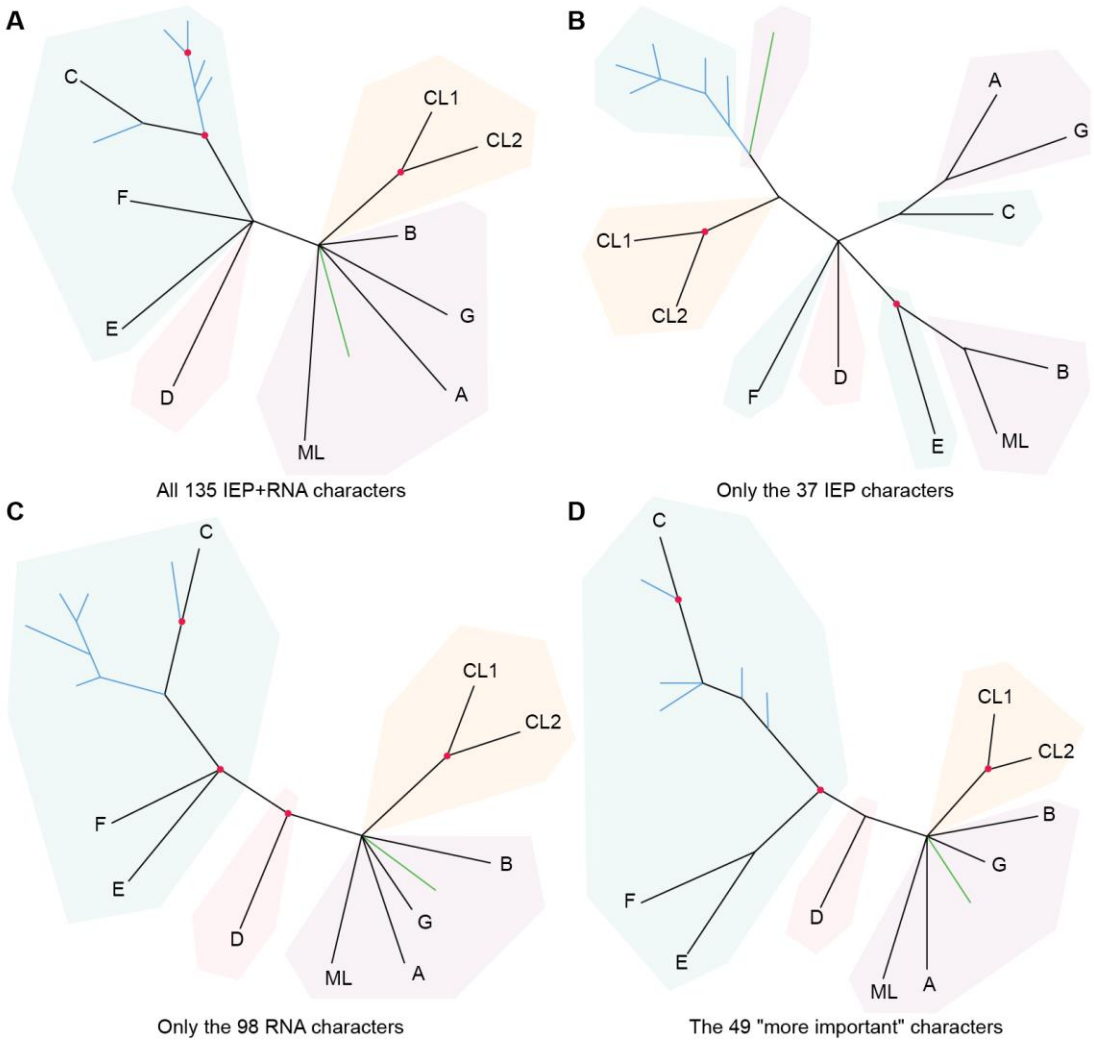


Figure 38. Morphological tree of group II introns.

Morphological data were extracted from both IEP and RNA at the basis of class. Four trees were constructed by MrBayes using **A)** all 135 characters from both IEP and RNA, **B)** only the 37 IEP characters, **C)** only the 98 RNA characters, and **D)** a subset of 49 characters that were considered being more important. Red dots indicate highly supported nodes (posterior probability >= 0.95). Blue lines indicate unclassified introns except Hp.au.I3 that is in green. Shadings in different colours correspond to the trend as proposed in main text. The complete information of all characters is available in Appendix D2.

Although it would be ideal to examine each individual intron, this is less feasible considering the amount of manual work. Possible improvements could be done next by including more representative sequences, such as at a basis of each subclade, for each class instead of treating one class as a single taxon. In addition, while using multiple sequences, it would also be possible to combine these morphological data together with regular sequence data.

Relationship with other RTs and the tentative origin of group II introns

Since the overall relationship between classes was usually consistent across multiple trees constructed in this project, the last analysis was to find the origin of group II introns. Group II introns have been hypothesized to be the ancestor of eukaryotic RT-containing elements such as non-LTR retrotransposons [58, 32], and thus the possible origin of group II introns could be predicted by rooting with other types of RTs. By taking advantage of sequence alignments prepared in Chapter II and other published literature, 31 types of non-group II RTs were used as external groups (Appendix D1). Initially, all external sequences were combined into one alignment with group II introns, but a preliminary NJ tree showed an extreme poor resolution while many members from different classes appeared to be mixed up (not provided). Therefore, each external RT group was individually aligned with group II introns with a specific mask in order to obtain a better tree resolution.

In order to reduce computational time, only a subset of 67 group II intron representatives were used together with each external type (Appendix D1). A tree including the 67 introns was first constructed, and showed consistent topology compared to the global IEP tree (Figure 39). Because a lack of alignable characters is common between group II and other types of RTs, ambiguity was expected and the relationship between group II introns and each non-group II RT class was approximately mapped onto this tree of 67 taxa (Figure 39). Over half (18 of 31) the external types, including non-LTRs, were most closely related to either Class C or the cluster of C+F+unclassified. These are more probably to be the earliest group II intron classes, which was also consistent with previous findings [264]. Meanwhile, the relationship between group II introns and some newly identified types, such as the unknown groups, were also estimated.

Although Classes C, F and unclassified introns could all be possible origins, Class C was expected to be the most probable if group II introns evolved parsimoniously. Class C introns have the smallest RNA ribozyme. While rooting group II introns at Class C, CL1+CL2 will be the newest branching classes. Introns of both CL1 and CL2 have the largest RNA on average, indicating the earliest form of intron RNA could be small, and gradually gained more structural features during evolution as a result of functional constraints or through recombination events. Accordingly, the IIB RNA structure (Classes B, D, E, F, G, CL1, CL2) possibly evolved directly from IIC (Class C); and IIA or IIA-like RNA structure (Class A, ML) originated from IIB, as could be seen that Class

tended to be close with ML in most trees inferred in this project.

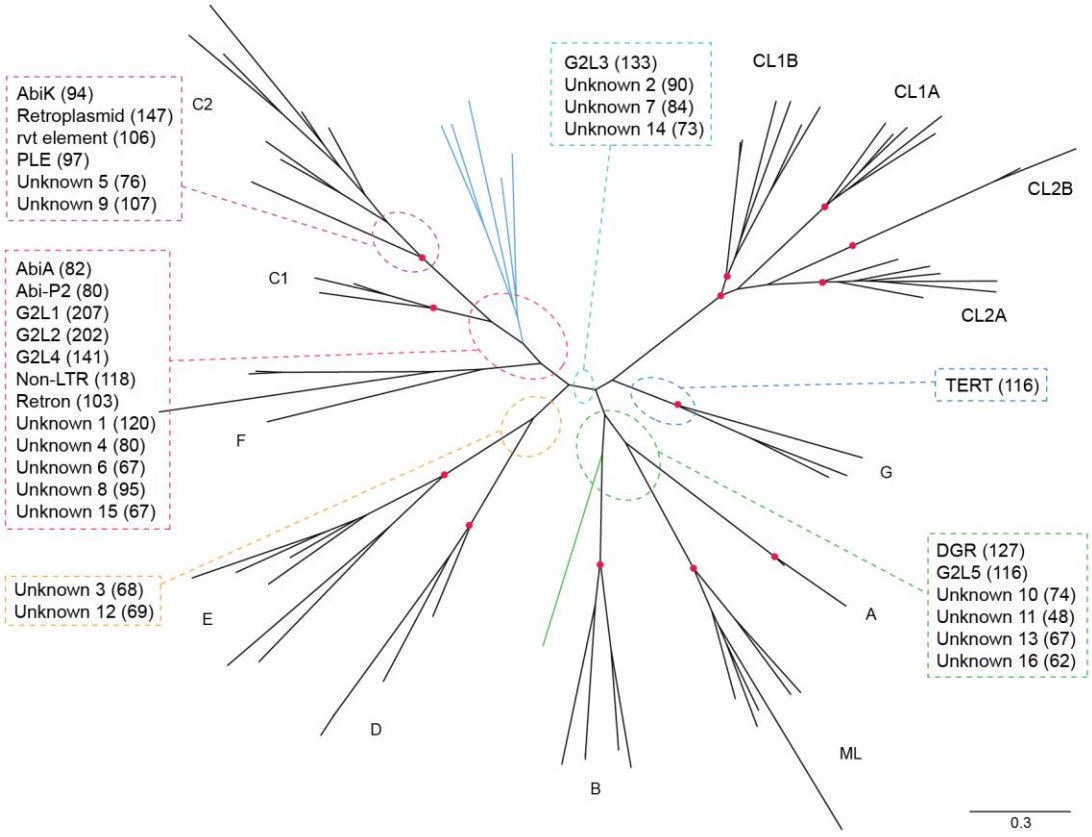


Figure 39. Relations between group II introns and external RTs.

Trees containing both group II introns and external RTs were constructed individually for each external group, and locations of external RTs are mapped onto the representative group II IEP phylogeny inferred using 67 introns. Nearby locations were combined. Numbers in parentheses indicate the number of amino acids used during tree inference. Red dot on tree indicate a highly supported (bootstrap >= 75%) clade at the class or subclass levels.

This parsimonious scenario also agrees with the IEP domain composition. As the only classes that contain an EN domain are B, CL1, CL2 and ML, which are less related to the tentative root of Class C. Therefore, the earliest form of the IEP (Class C) probably did not contain the EN domain, which was gained later by classes evolved later. Based on the sequence, it was also noticed the EN domain of CL1 and CL2 is less similar to that of Class B and ML, indicating the process of gaining the EN domain occurred independently in different classes.

Conclusion

This project aimed to illustrate the relationship between classes of group II introns and to depict their evolutionary histories using a large dataset. Similar to the previous study in 2009 [181], phylogenetic trees were inferred using both the IEP and RNA sequences. Multiple sequence masks were compared, and the larger masks tended to produce trees with higher resolution but overall there were minimal differences between various masks.

Generally, both the IEP- and RNA-based trees showed similar topologies, even though the RNA trees were less robust, probably due to the lack of informative characters. However, a few disagreements at the class level were observed while comparing the IEP and RNA trees: the disagreement involving CL1, CL2 and their subclasses were solved by CL-specific trees; the disagreement involving Classes B and G was rather of interest as it could indicate a swapping of the IEP gene; while other disagreements remained unsupported, even though they have been observed in multiple trees.

As mentioned previously in the 2009 study [181], trees of group II introns seemed to be sensitive to taxon sampling and base composition heterogeneity. Neither of the IEP- and RNA-based trees were sensitive to taxon sampling this time, possibly because the larger dataset enabled a more even selection for subsets of introns. In contrast, the base composition heterogeneity still seemed to affect the tree. The larger dataset, which was likely a better representation of the entire intron population in nature, could diminish this issue.

Based on the topologies, the IEP and RNA have possibly coevolved. However, when trees were subjected to the more formal SH topology tests, almost all IEP trees were predicted to be conflicting with the RNA trees. However, through a more detailed examination using Class A as an example, it was eventually implied that trees of group II introns are not suitable for topology tests, as they are too divergent in sequence and are expected to frequently reveal alternative topologies. Although such alternatives are more often minor disagreements and would not affect the interpretation of the evolution of group II introns, they could cause the SH test to reject the null hypothesis. As a different attempt, trees were collapsed when node supports were less than 75%, however, even more of such collapsed trees were rejected by the SH test (data not shown). Therefore, the results of SH tests were dismissed, while all trees were then compared manually, which resulted the same conclusion that most introns (if not all) had their IEP and RNA

coevolved.

Next, as an alternative approach to restore dismissed information in both of the IEP and RNA alignments due to the lack of sequence similarity, structural features were selected for each class and coded as morphological data, which was used to construct morphological trees. These trees generally agreed with the sequence-based topology. They can be used as additional evidence for the often observed yet unsupported relationships between classes. Finally, group II intron IEPs were aligned with other types of RTs in order to identify their origin. While both Classes C and F, together with unclassified introns, could be the origin, Class C had the highest probability if assuming a parsimonious evolutionary history.

Overall, this project used a large dataset and constructed trees that showed agreeing topologies, although often without supports. In addition to repeating analyses done in the 2009 study using a larger dataset, this project included new analyses, such as using various masks, constructing class-specific trees and morphological trees. To construct better morphological trees could be the next point. Criteria used to define and classify structural features need to be first refined, and ideally automated; next, more taxa could be introduced into the morphological dataset, and gradually integrated with the sequence-based data.

Chapter V. A manually curated comprehensive DGR compilation

Abstract

Diversity-generating retroelements (DGRs) are retroelements that function to diversify the protein sequence of their target genes (TGs). While the three DGRs from *Bordetella* phage, *Treponema* and *Legionella* have been well characterised experimentally and showed variations in both function and gene structure, more DGRs have been predicted from genomic data and have shown even more variation. To investigate the diversity of DGRs, this project compiled a set of 372 DGRs using an automated prediction and followed by manual curations through various analyses. The analysis showed this compilation not only included an expanded set of DGRs that are similar to that of the *Bordetella* phage, but also discovered a large number of novel DGRs from the CPR (candidate phylum radiation) group. Each component was systemically examined and all DGRs were classified based on either the variable region (VR), the domain composition of the TG, or the entire DGR cassette structure, which revealed a great diversity of DGRs in all these aspects. The classification defined two new accessory genes, MSL and CH1. Next, the TR-VR pair of each DGR was examined and showed it could be common for DGRs to introduce other substitutions aside from the canonical A-to-N substitution to expand the possibilities during TG diversification. In addition, the association of DGRs to phages was investigated, and revealed that both bacterial hosts and phage hosts are common for DGRs. Finally, most DGR components were noticed to have a correlation with the RT-based phylogenetic tree, indicating DGRs were more likely evolving as a single unit rather than having diverse histories for each component. This compilation provided a comprehensive list of DGRs that were manually curated and can be used as references to direct future DGR-related research.

Introduction

DGRs are retroelements consisting of multiple gene components and have beneficial functions to the host. They are found mainly in phages, bacterial genomes and plasmids [19, 62, 161]. Different DGRs vary in their components and organisation, but they all have three essential components: a reverse transcriptase (RT), a template repeat (TR), and a target gene (TG) that often contains a variable region (VR) at the 3' end (Figure 8, Figure 40). They benefit the host cell through mutagenic retrohoming, in which the RT reverse transcribes the TR to produce a piece of DNA with random A-to-N mutations. This DNA copy of TR, which is slightly different in sequence compared to the original TR, replaces the VR of the TG at the C terminus. As a result, the TG will have a different amino acid sequence after acquiring the new VR sequence. Most identified TGs are involved in cell surface display, and the DGRs increases the adaptability of their hosts (Figure 8) [19, 20, 62, 160].

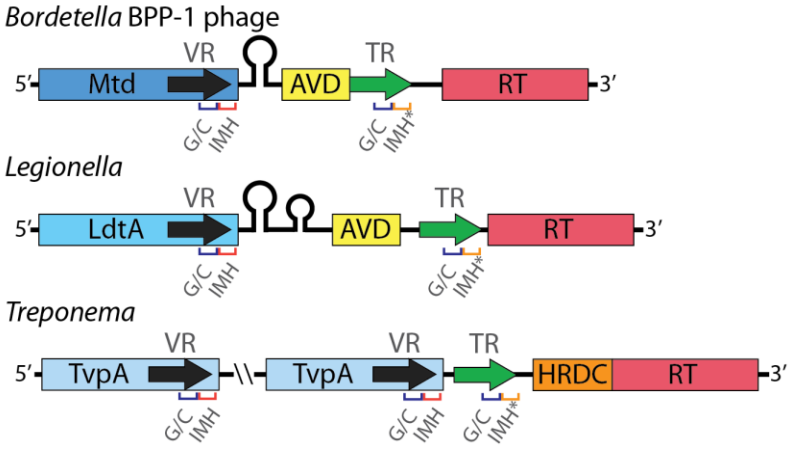


Figure 40. Comparison of three experimentally characterised DGRs.
The three DGRs are from *Bordetella* BPP-1 phage, *Legionella* and *Treponema*, respectively. They vary in gene organisation, the type and number of the TG, and the type of the accessory gene. Elements are not drawn to scale. Figure adapted from references [19, 169, 161, 291].

The best experimentally studied DGR is from the *Bordetella* BPP-1. It targets the phage *mtd* gene, which is responsible for cell binding during phage infection (Figure 40) [62]. Crystal structures of the Mtd protein showed that its C terminus, including the VR, has a CLec (**C**-type **lec**tin) fold, which can tolerate massive sequence variation to ensure binding diversity [167]. The VR also encodes a few constant amino acids, which serve

as a structural scaffold for the protein [167]. In addition, this DGR contains an accessory gene *avd* (**a**ccessory **v**ariability **d**eterminant) (Figure 40), which plays an essential role during cDNA synthesis by binding to both the RT and cDNA [162, 292].

Using the *Bordetella* phage DGR, it has been demonstrated that the recognition between the TR and VR as well as directional homing require multiple elements. At the 3' end of the VR of the *Bordetella* phage DGR, there is a 14 bp GC-only element (usually referred as "G/C" or "$(G/C)_{14}$") followed by an IMH (**i**nitiation of **m**utagenic **h**oming) element. Further downstream, there are two GC-rich inverted repeats that can form a DNA stem-loop structure (Figure 40) [62, 164, 293, 294]. Portions equivalent to both G/C and IMH elements can also be found in the TR. The G/C in TR has an identical sequence compared to that of the VR, while the IMH in TR differs from the VR in sequence and is thus named IMH* (Figure 40) [294]. Unlike the VR, there are no inverted repeats downstream of the TR. All of the G/C, IMH, IMH*, inverted repeats, along with the sequence similarity between VR and TR ensure precise VR recognition and directional homing: the recognition at the 5' end of the VR is dependent on the sequence similarity between TR and VR, while the recognition at the 3' end of VR is independent from the sequence similarity and requires structural features from other elements [62, 292, 293, 294].

Besides the DGR from the *Bordetella* phage, there are also other experimentally characterised DGRs, such as those from *Legionella pneumophila* [169] and *Treponema denticola* [170]. The *Legionella* DGR has a gene structure in the order of TG-AVD-TR-RT, which is similar to the *Bordetella* phage DGR (Figure 40). It targets the *ldtA* (***l****egionella* **d**eterminant **t**arget **A**) gene, which encodes a lipoprotein located in the outer face of the outer membrane that is involved in ligand binding [169]. Unlike the *Bordetella* phage DGR, the *Legionella* DGR has two pairs of inverted repeats downstream of the VR, and form two stem-loop structures (Figure 40) [169].

In contrast to both *Bordetella* phage and *Legionella* DGRs, the *Treponema* DGR has an order of TG-TR-HRDC-RT, in which HRDC (**h**elicase and **R**Nase**D** **C**-terminal) is the accessory gene instead of the AVD (Figure 40). It targets the *tvpA* (***T****reponema* **v**ariable **p**rotein **A**) gene that is also thought to have a binding function involved in interacting with other cells [170]. Moreover, the *Treponema* DGR has been found to target multiple TGs located distant from the DGR cassette (Figure 40) [170]. For both *Legionella* and *Treponema* DGRs, all G/C, IMH and IMH* elements could be identified (Figure 40) [169,

170]. In addition, both TGs have been predicted to have a CLec fold around the VR, even though their VR sequences have a low sequence identity compared to the VR of the *Bordetella* phage DGR, indicating bacterial DGRs often target proteins involved in surface display and ligand-binding [169, 170].

Because of the characteristic A-to-N mismatch between the TR and VR, putative DGRs can be predicted directly from genomic sequence data [19, 293, 160, 295, 296, 297, 291, 298]. Many past searches have identified DGRs that vary in the composition and order of the gene components, as well as the properties and number of each different component [19, 160, 293, 298]. More recent studies also identified novel DGRs from metagenomic data of the DPANN (**D**iapherotrites, **P**arvarchaeota, **A**enigmarchaeota, **N**anoarchaeota, **N**anohaloarchaea) and CPR (**c**andidate **p**hylum **r**adiation) groups [296, 299], further indicating the widespread existence and great diversity of DGRs.

This project aimed to investigate the diversity of DGRs and began to collect a set of DGR representatives from the most recent version of the GenBank database through bioinformatic approaches. Followed by manual correction, a total of 372 DGRs that are unique in sequence were finally compiled and subjected in various studies, such as the classification for each gene component and phylogenetic analyses. This manually curated set of DGRs can be used as reference for future research.

Materials and Methods

An updated identification of primary DGR components

Before I started working on this project, an unpublished set of ~500 DGRs was collected by Michael Abebe when he worked in our lab in 2012. He performed several preliminary analyses for those DGRs, including identification of all DGR components, the copy numbers of each component, the order of genes and tentative classifications for the TG. In this section, they are referred as the "old" dataset.

Since I began this project in 2013, I either made updates or deleted some DGRs from the old dataset. Because of the discovery of the new and diverse CPR group in 2015 [300, 301], and also in order to look for novel DGRs from the most recent version of the GenBank database, I performed another search in 2016 and discovered ~130 additional putative DGRs. This section will describe how new DGRs were discovered, which mainly

included searching of the RT, the TR-VR pair, the TG, and accessory genes.

Identification of DGR-related RTs

The first step to identifying a DGR is to search for DGR-related RTs. This was done using a BLAST-based strategy similar to that described in Chapter II. First, 55 representative RTs selected from the old set were used as query sequences (Appendix E1), and BLASTP (protein queries against the protein database) searches were performed against GenBank's non-redundant protein database (NR) as of July 25, 2016.

All resulting hits were subjected to another BLASTP search in order to filter out those that are unlikely to be related to DGRs. For the second BLASTP search, each of the hits was used as the query and searched against a custom RT database, which consisted of RTs from different classes (Appendix A2). If at least one of the top three hits was not a DGR-related RT, the query was considered as a non-DGR RT and removed.

Finally, the nucleotide sequences of the remainder of the RTs were downloaded with ± 10 kb flanking each side for later analysis. For convenience, they will be referred to as the "downloaded sequences" in the following text.

Because the search I performed in 2016 mainly aimed to find novel DGRs rather than to create a comprehensive set of DGRs, DGRs that were similar to the old set were excluded. This was based on the similarity of the RT. First, all newly discovered candidate RTs were compared to RTs from the old set using BLASTP. An RT would be removed if it had an E value of less than 1e-100 and if the hit coverage was larger than 85% of an RT from the old set. Next, the remaining RTs were automatically aligned to those from the old set using the program MUSCLE [213], and the alignment was used to generate a neighbour-joining (NJ) tree using the program Clustal X [211] with default settings (tree not provided). Sequences that were grouped to RTs from the old set were considered similar and therefore excluded. Meanwhile, in the case of many RT sequences being grouped together and supported in a single clade, but with extremely short branch lengths, only one of them was retained while the rest were considered duplicates.

Identification of TR-VR pairs, TGs and VR-based classification

The TR-VR pairs were identified from the downloaded sequences by searching for short sequence repeats specifically containing A-to-N mismatches using BLASTN. The BLAST algorithm was set to "blastn" instead of the default "megablast" in order to allow short repeats to be matched. Each downloaded sequence was used as the query and searched against itself. Results were evaluated as "high", "medium" or "low" based on criteria considering the length of match and the ratio of A-to-N mismatches (Table 6). Only those evaluated as either "high" or "medium" were retained for the following steps.

Table 6. Criteria for evaluating the TR-VR pairs.

Many factors, including the length, number of all mismatches, and the proportion A-to-N mismatches, were considered and each potential TR-VR pair was put into one of three categories: "high", "medium" and "low", which correspond to the possibility to be associated to a real DGR cassette.

	High	Medium	Low
Length	>= 50 nt	>= 40 nt	>= 40nt
Mismatches	>= 8	>= 6	>= 6
Proportion of A-to-N vs. all mismatches	>= 80%	>= 70%	>= 50%

After determining the TR and VR, the longest ORF containing the VR at the C terminus was predicted as the TG. Other ORFs were considered if the VR was not located at the C terminus. Entries in which the TG could not be predicted were excluded from the dataset.

VRs were then classified based on the sequence similarity. Prior to my project, five unpublished VR classes, CLec1-3 and Ig1-2, had already been established in the old set. Therefore, their sequence alignments were used to create profiles with the program HMMER [214]. The newly identified VRs were compared to these profiles, and those that showed matches were assigned to the same class. TGs that matched none of these profiles were then automatically aligned using MUSCLE and classified based on the sequence similarity.

Identification of accessory genes

Accessory genes were identified using either HMMER-based profile search or pairwise BLASTP comparison. The HMMER-based search aimed to find known accessory genes. Prior to my project, the old set contained four types of accessory genes: AVD, HRDC, MSL ("MutS-like", see Results section) and CH1 (conserved hypothetical). Similar to the classification of VRs, the program HMMER was used to create profiles of these four types from their sequence alignments, and these profiles were then used to identify accessory genes for newly discovered DGRs by searching all ORFs that were longer than 50 aa (obtained through six-frame translation) from each downloaded sequence.

New types of accessory genes were classified using pairwise BLASTP comparisons. All ORFs that are less than 1 kb from other identified DGR components (RT, TR, TG) were subjected to this comparison. Every two ORFs were compared using BLASTP, and grouped based on the E value. Only groups with three or more members were retained and assigned as **p**otential **a**ccessory **g**enes (PAGs).

Summary of the final dataset

The final dataset did not contain DGR duplicates based on the RT sequence identity. If multiple DGRs had an RT that was >= 95% identical in the amino acid sequence to each other, only one DGR was retained in the final dataset as a representative. Combing both the old dataset and DGRs I discovered in 2016, a total of 368 "unique" DGRs were included in the final set. In addition, four published DGRs found in *Nanoarchaeota archaeon* [296] were appended into the final dataset. They were missed by the automatic search because they are from the WGS (whole genome shotgun) database, which was not searched by the automatic method. Therefore, the final set contained 372 DGRs.

Based on a phylogenetic tree using the RT sequence (see later), 127 DGRs from CPR phyla together with the four nanoarchaeal DGRs are put into one subset named "CPR", and the rest of the 241 DGRs were put into a subset named "core".

Analyses applied to all 372 DGRs in the final set

Phylogenetic inference using DGR-RT sequences

Although a NJ tree was generated earlier during the identification of new DGRs, the NJ tree was not expected to be able to rigorously reflect the relationships among all RTs. Therefore, maximum likelihood estimation was used to construct another phylogenetic tree of all RTs from the final dataset. The RT sequence alignment was first generated by HMMER using a profile created from the RT alignment of the old dataset, followed by manual refinements. Of this RT alignment, 149 characters that were considered as being more conserved were used to construct the tree using the program RAxML [226] with 1000 bootstrap replicates. Settings used for RAxML are as: the substitution matrix was RtREV [302], the proportion of invariable sites estimation was considered, and the heterogeneity of substitution rates was set to gamma distribution (command was written as "-m PROTGAMMAIRTREV" for RAxML). Other settings were used as default. The program was run on the Bugaboo server of WestGrid (https://www.westgrid.ca).

Identification of remote VR, TG and accessory genes

Remote VRs (outside of the downloaded ~20 kb sequence) were identified by using previously determined TR sequences as the query and searching against the complete genomic or contig sequences retrieved from GenBank using BLASTN. All remote VRs were verified that they do not belong to a different DGR cassette. Accordingly, the longest ORF containing each remote VR at the C-terminus was assigned to be the remote TG.

Remote accessory genes were identified through HMMER-based profile searches for all currently known types: AVD, HRDC, MSL, CH1, and PAG1-4. The search was also performed against the complete GenBank record for each DGR.

Analysis of protein domain composition and protein fold of the TG

The domain composition for both the TG and RT was analysed by CDD (Conserved Domain Database, https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) [236] through the online batch search interface. All TG and RT sequences were submitted, and the

output was refined using custom Perl scripts (not provided). First, weak matches with an E value greater than 1e-5 were excluded. Next, overlapping hits belonging to the same superfamily were merged to reduce redundancy. In addition, based on CDD annotations, many superfamilies were combined because they share some degree of sequences similarity and are believed to have the same function. This project combined two sets of superfamilies. The first one included superfamilies "AAA_16" and "P-loop_NTPase", which are all referred to "P-loop_NTPase"; the second one included superfamilies "DUF823", "Fib_succ_major" and "DUF1566", which are all referred to "DUF1566".

The protein fold of each VR sequence was predicted by the Phyre2 webserver (http://www.sbg.bio.ic.ac.uk/~phyre/) [220].

Identification of the GC-rich inverted repeats

Because the GC-rich inverted repeats are downstream of the VR and often form stem-loops, the program Mfold [218] was used to fold automatically the ~200 bp of DNA downstream of the VR. The type of molecule was specified as DNA and the rest of settings were all used as default. The folded DNA structures were output in the Vienna format, and were screened with a custom Perl script (not provided) to search for GC-rich stem-loops using criteria considering the size, GC content, mismatches and the distance to VR (Table 7).

Table 7. Criteria for identifying GC-rich stem-loops downstream of the VR.
Potential stem-loop structures were divided into two categories with high and low probability of being a real GC-rich stem-loop. Structures that do not meet either category were considered to have no probability and excluded.

	High	Low
Stem length	>= 5 nt	
Loop length	<= 6 nt	
G-C pairs in stem	>= 50%	
G-T pairs in stem	<= 20%	
Bulge/mismatch	0	<=2
Distance to the end of VR	<=50 nt	<=100 nt

Detection of transcriptional terminators downstream of the VR

Putative transcriptional terminators downstream of the VR were detected using the web server ARNold (http://rna.igmors.u-psud.fr/toolbox/arnold/index.php). The input was the ~200 bp DNA sequence downstream of the VR, both strands were analysed, and the rest settings were used as defaulted by the web server.

Estimation of phage-association for DGRs

Three approaches were used to judge whether a DGR could be associated with a phage. The first was based on the GenBank annotation, and DGRs from a phage source were automatically considered to be associated to a phage. The second approach considered DGRs targeting a Mtd-related protein as being associated to a phage, as Mtd is the TG of the *Bordetella* phage DGR. The last approach was based on BLASTP searches against a database available on the PHASTER website (http://phaster.ca/databases) [303], which consisted of only phage-related proteins. All ORFs of the downloaded sequences were subjected to this search. Only hits to a phage structural protein with an E value less than 1e-5 were considered, including baseplate, capsid, collar, head, neck, portal, sheath, tail, tape measure, terminase and whisker proteins. The final evaluation was based on the number of matches at each side of the DGR.

Results and Discussion

An overview of the DGR compilation

Through both automatic prediction and manual curation, this project aimed to collect and analyse a comprehensive set of DGR representatives from the GenBank database. The automated identification was derived from a previously published pipeline program for identifying group II introns [245]. It began with identification of DGR-related RTs, followed by searching for TR-VR pairs that have the characteristic A-to-N substitutions. Next, TGs that contained the VR at the 3' end were confirmed, and accessory genes were identified by using either known sequence profiles or by looking for similar protein groups. Two searches were done to collect DGRs. The first search was performed by Michael Abebe in 2012, and in 2016, I performed the second search to look for novel

DGRs from newly published sequences.

All candidate DGRs were initially kept in the dataset. In order to reduce the time for manual analysis, DGRs that have the RT >=95% identical in sequence to each other were reduced to one representative. It was noticed that many DGRs had duplicates in either the same or a different host organism, and the total number of DGRs would increase about 3-fold if duplicates were included.

Because this project only focused on putatively functional DGRs, defective DGRs that either lacked a core component (RT, TR, VR) or had apparently non-functional components were excluded. However, it was also noticed that there is a rather large number of inactive DGRs compared to putatively functional "full-length" DGRs, implying DGRs are frequently transferred across individuals.

The final dataset contains a total of 372 unique DGRs. Of them, 368 were automatically identified through the two searches. The remaining four DGRs are from *Nanoarchaeota archaeon* [296] and were appended into the final dataset. They were missed by the automatic search because they belong to the WGS database, while the automated method was designed to only search for the NR database.

Based on the host organism, the final dataset was divided into two subsets, named "core" and "CPR (**c**andidate **p**hyla **r**adiation)" respectively. The core set contains 246 DGRs that are mostly from Bacteroidetes, Cyanobacteria, Firmicutes and Proteobacteria, which are more commonly sequenced phyla. In contrast, the CPR set contains 126 DGRs, 122 of which are from the bacterial CPR group [300, 301]. The remaining four are from the archaeal DPANN (**D**iapherotrites, **P**arvarchaeota, **A**enigmarchaeota, **N**anoarchaeota, **N**anohaloarchaea) group [296].

Both CPR and DPANN groups were discovered from metagenomic or single-cell sequencing data [300, 304]. Species belonging to either group are generally small in genome size and have restricted metabolic capacities related to carbon- and nitrogen-cycling, and are thus thought to have a symbiotic or parasitic life cycle [300, 301, 305, 304, 306]. While both groups are diverse, the CPR group has been expected to contain over 70 candidate bacterial phyla that may occupy more than 15% of all bacterial diversity [301]. Accordingly, DGRs from these species have been found to be separated from those of the core set in terms of the phylogeny, the gene order and the VR-based classification. Therefore, the CPR set indeed differs from the core set not only because

of the host organisms but also because of many features of the DGRs themselves. Note that even though CPR and DPANN do not have a direct biological relationship to each other, DGRs from both groups were merged and only referred to the CPR subset due to the DPANN group was new to the "old set", and they had a small group size of 4.

Overall, this compilation is believed to include a comprehensive set of DGRs from the most recent GenBank database. In the following sections, all information gathered through this project will be described, and summaries of all 372 DGRs are listed in a "master table", which is provided in Appendix E2.

Establishment of an RT-based phylogenetic tree of DGRs

In order to illustrate the relationships among all DGRs, the first analysis was to construct a phylogenetic tree using the RT sequences. Although a DGR contains multiple components, only the RT is present in all cassettes and all RTs share sequence similarity. The tree was constructed by maximum likelihood estimation using the program RAxML with 1000 bootstrap replicates. As shown in Figure 41, this tree agrees with the division between the core and CPR (highlighted in pale yellow sectors) subsets, in which almost all DGRs from the CPR set belong to a large single clade and are separated from the core set (a few exceptions belong to the DPANN group). For the core set, there are a few large clades that can be observed, but almost all DGRs from the CPR set do not form large clades at all (Figure 41). However, including the separation of the CPR subset, this tree generally lacks supports (bootstrap >= 75%) for most of the clades and the resolution is rather low.

Regardless of the lack of supports, the topology of this tree agrees with an earlier preliminary tree constructed by Bayesian inference using a smaller dataset (data not shown), and thus major clades observed in this tree were believed to represent different lineages of DGRs. Therefore, based on both the tree topology and many other features such as the VR-based class and gene order (later sections), four lineages (1-4) were assigned to four major clades (Figure 41, grey sectors). Within lineage 3, a subclade was further assigned as lineage 3a (Figure 41, darker grey sector) because all members have the same gene organisation, which is different from other DGRs from lineage 3 (later section).

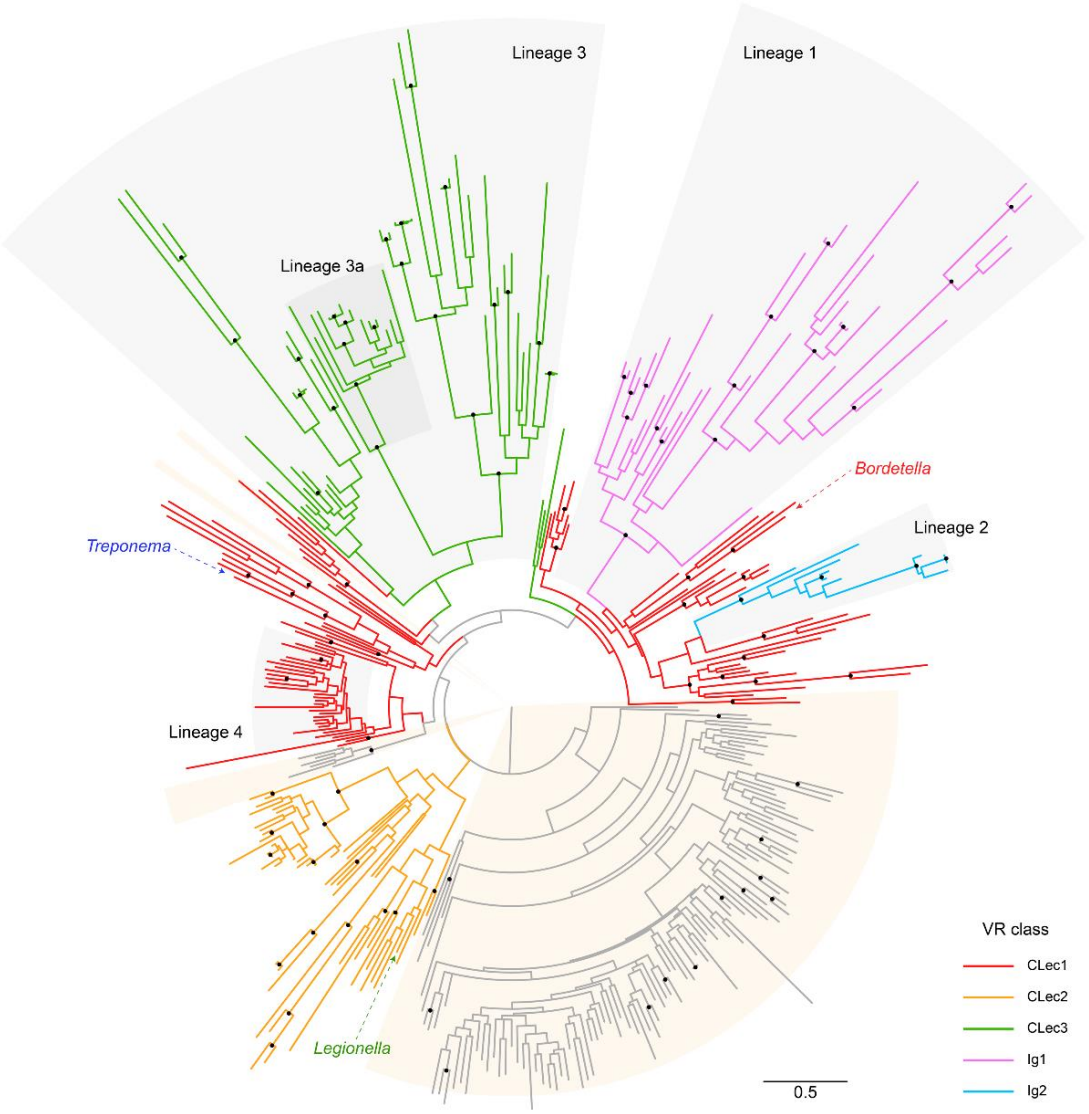


Figure 41. Phylogenetic tree of DGRs based on the RT sequence.

The tree was generated by program RAxML with 1000 bootstrap replicates. The input sequence data contained 149 characters that were considered to be the most conserved. Clades with a >= 75% bootstrap support are marked by a black dot. Grey sectors correspond to Lineages 1-4, and the darker grey sector corresponds to Lineage 3a. Pale yellow sectors correspond to DGRs from the CPR set. The three characterised DGRs from the *Bordetella* phage, *Legionella* and *Treponema* are highlighted by dashed arrows in red, green and blue respectively. Branches are colour coded by major VR classes (CLec1-3 and Ig 1-2).

Components of the DGR cassettes

Reverse transcriptase

The RT gene is the only component present in all DGRs that shows sequence similarity. Therefore, the RT sequence could be used as the first step to search for new DGRs. Overall, RTs from this compilation have the same organisation of RT motifs 1-7 and an X domain, which are alignable to RTs from group II introns. However, unlike group II introns, DGR-RTs do not have a nuclease domain [22, 62, 296] (Figure 13 of Chapter II). Even though a few RTs have a C-terminal extension of ~150 aa, which might correspond to a nuclease domain, there was little evidence that this extension had a nuclease motif or a related activity.

Through CDD-based domain identification, most RTs did not show additional motifs except for three, which showed a weak match to a MutS_I motif that is involved in repairing mismatched DNA. Because other MutS motifs have been identified in some accessory genes (see later section), this implies that some DGRs may use components related to the DNA repair system during mutagenic retrohoming, which is different from the *Bordetella* phage DGR prototype [164].

Template repeats and variable regions

The TR-VR pairs were detected by searching for sequence repeats with A-to-N mutagenesis. Unlike the RT, not all VRs can be aligned, and sequence similarity was only seen in subsets of VRs. Therefore, VRs are classified into five major and seven minor groups based on the sequence similarity (Figure 42). The five major classes include CLec1-3 and Ig1-2, in which "CLec" and "Ig" are unrelated protein motifs, and they correspond to "**C**-type **lec**tin" and "**i**mmuno**g**lobulin" folds respectively. TGs from these major classes were all predicted to have a corresponding protein fold by the Phyre2 webserver.

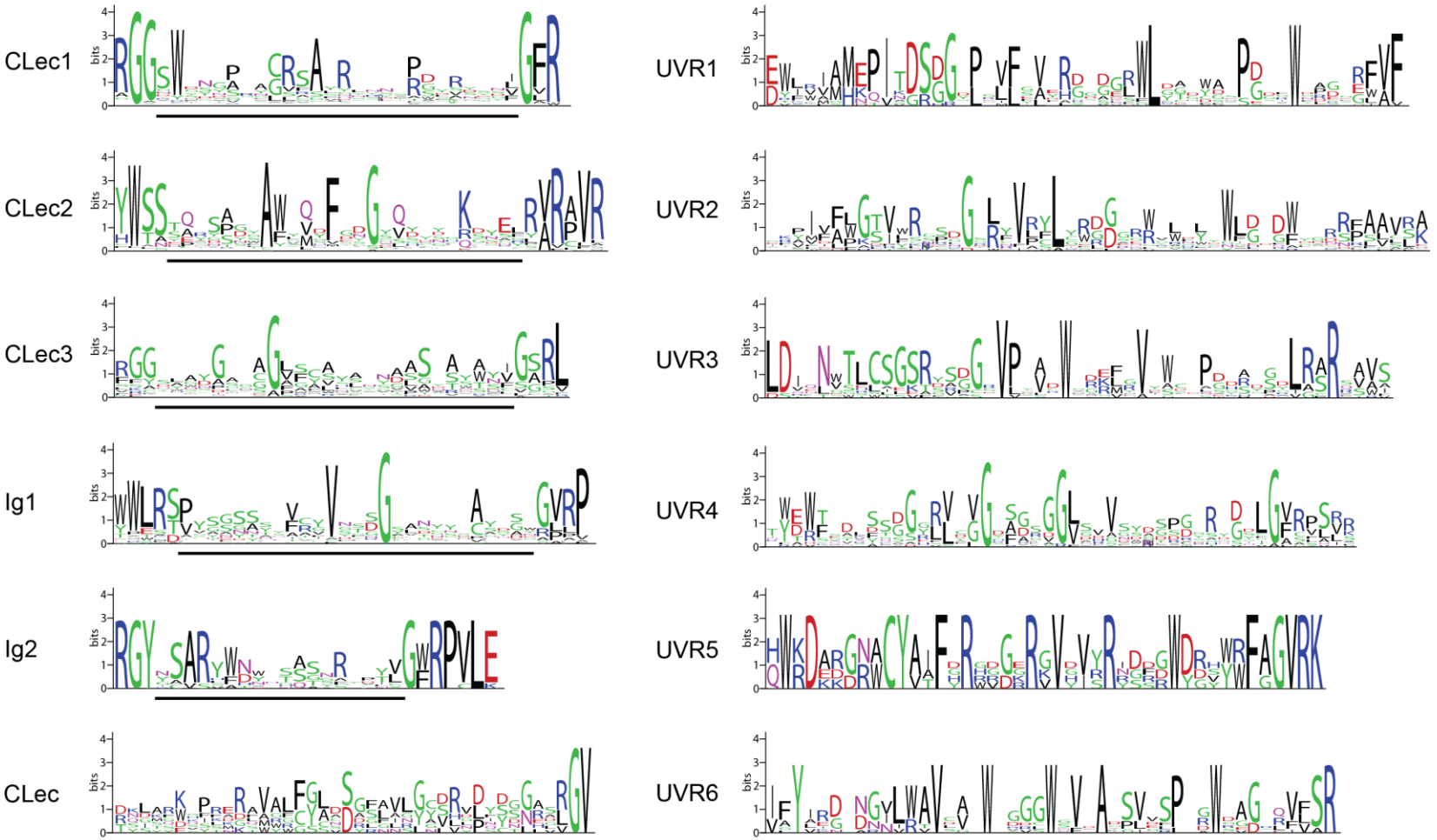


Figure 42. The WebLogo profile of each VR class.

Images were generated by the WebLogo server (http://weblogo.berkeley.edu/logo.cgi) [217] using the VR alignment of each class. For the five major classes (CLec1-3 and Ig1-2), a thick line is placed below areas subjected to the A-to-N mutagenesis.

The seven minor classes include CLec and UVR1-6, in which "UVR" stands for "**u**nknown **VR**". Members of the minor CLec group could be identified to have a CLec fold, but they could not be aligned to any of the major classes of CLec 1-3. None of the six UVR groups were predicted to have a CLec or an Ig fold, nor could they be predicted to have an alternative protein fold. Regardless, all UVR classes have conserved amino acids at both ends as well as in the middle of the VR (Figure 42). This is consistent with the pattern observed in CLec and Ig classes [167], and these conserved amino acids are likely responsible for maintaining a scaffold to support the display of the diversified amino acids at the surface of protein.

A clear correlation between the five major VR classes and the RT-based tree was observed (Figure 41), indicating the VRs were evolving together with RTs. While CLec3, Ig1 and Ig2 are associated to Lineages 1, 2 and 3 respectively, CLec2 correlates within one single clade (Figure 41). In contrast, CLec1 is found in multiple clades, suggesting CLec1 VRs could be the oldest. In contrast, the six UVRs were dispersed and showed no strong correlation to the tree, except that they all belong to the CPR set (Appendix F1). Not only because 45 VRs from the CPR set remained unclassified, but also because the CPR set did not form into large clades as shown in the RT-based tree, their VRs and TGs are might thus have different functions compared to those of the core set.

Inferred patterns of mutagenesis in VR sequences

During the step of detecting TR-VR pairs, one important criterion was the ratio between the A-to-N mutagenesis and all mismatches (Table 6). It was initially expected that a DGR would only contain the canonical A-to-N mismatches between the TR and VR, but this compilation revealed many other types of differences. By questioning whether it is common for a TR-VR pair to allow non-A-to-N mismatches, the TR-VR alignments (Appendix E3) were more closely examined.

The majority of DGRs in this compilation (65%, 243 of 372) were found to contain only A-to-N mutations, which mostly occur at the first and second codon positions (Figure 43A, B). This agrees with the previous conclusion that substitutions at the first two codon positions could allow the maximal possibility to diversify the amino acid sequence [163, 167]. In addition, when a DGR interacts with multiple TGs, they always result in non-identical VR sequences (Figure 43B), indicating the process of mutagenesis always

occurs randomly.

The remainder of the 35% DGRs in this compilation contain non-A-to-N substitutions and/or indels. Non-A-to-N substitutions (Figure 43D, E) were probably caused by random mutations. Non-A-to-N substitutions were observed at a low frequency. Only one such occurrence was the mostly observed (in ~55 DGRs), ~10 DGRs showed two occurrences, ~5 DGRs showed more than two, and one DGR was the most extreme case to have seven. Indels (Figure 43C, D, E) were even less frequent than non-A-to-N substations, and were mostly observed to occur next to the "AAC" repeats, which is a favoured "codon" and usually appears in tandem in TR to maximises the sequence variations after diversifying [19, 163]. Therefore, they were likely caused by template slippage during reverse transcription. They were observed the least frequently in this compilation, as only ~20 TR-VR pairs contain indels. Interestingly, the number of indels causing frameshifts is only slightly lower than those not causing frameshifts, while the IMH usually remained intact (Figure 43C, D, E).

Aside from the type of mutagenesis, the number of substitutions also vary in different TR-VR pairs. From this compilation, the number of A's in TR vary from 9 to 56, while most TRs contain 21-35 A's. Accordingly, 11-20 nucleotide substitutions are the most common between TR-VR pairs, and 21-35 nucleotide substitutions are less common, while more extreme examples have fewer than 10 nucleotide substitutions. In terms of the number of changes in the amino acid sequences, 6-15 amino acids are the most common, followed by 16-25, and only a few have fewer than 5 amino acid substitutions.

Overall, this showed that between the TR and VR, the type and number of mutations vary among different individuals. Although canonical A-to-N mismatches are the most favoured, non-canonical substitutions and indels are not rare. They were likely caused by mistakes during reverse transcription, but these errors may still increase the TG diversity if they do not deactivate the TG.

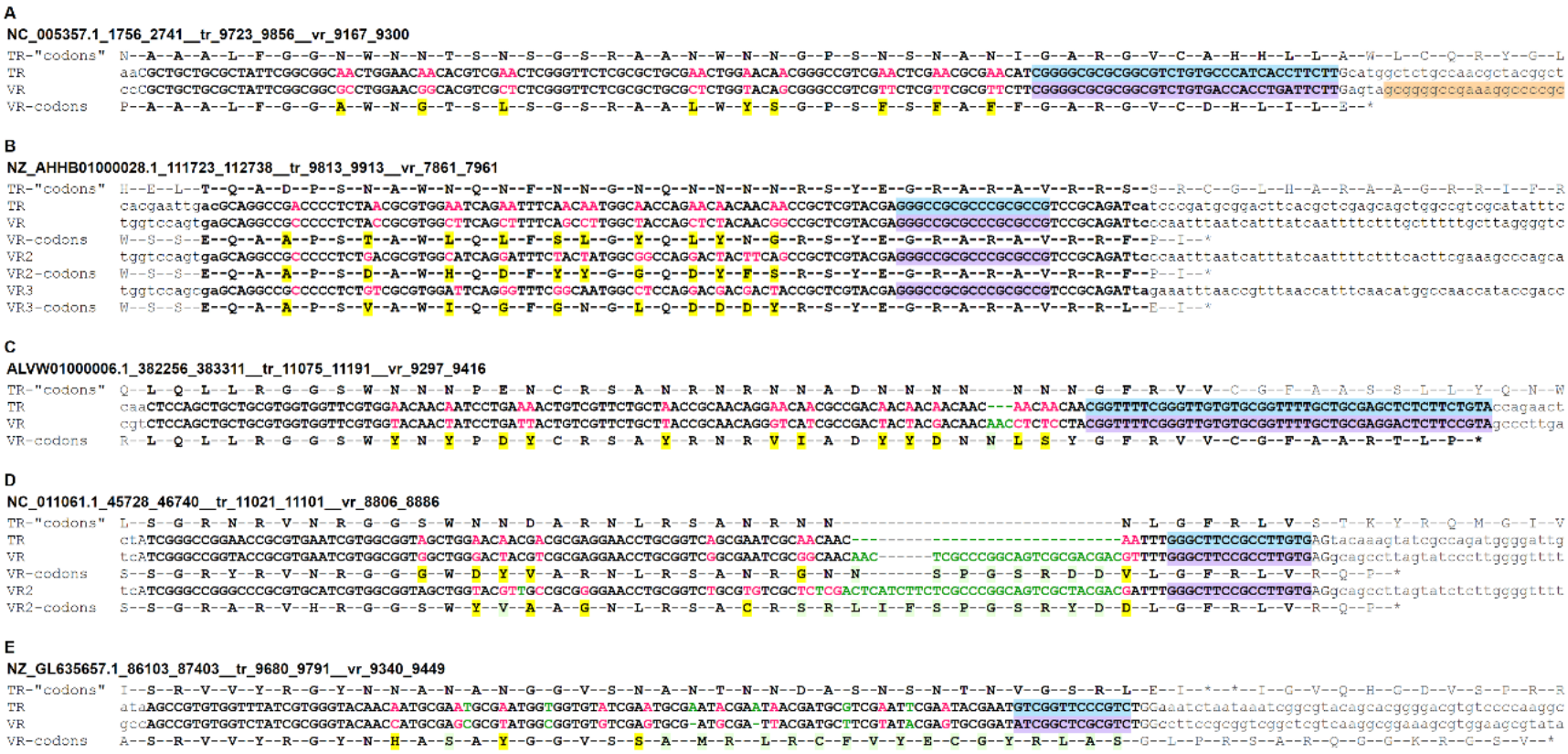


Figure 43. Selected TR-VR pairing examples.

Predicted TR and VR sequences are written in capital letters and alignable areas are in bold. Red indicates A-to-N mismatches, while green indicates non-A-to-N mismatches as well as indels. Changes of amino acids in corresponding translated VR are thus highlighted using yellow and green respectively. Blue and purple shadings corresponding to detectable G/C and IMH*/IMH elements. Orange shading in A) highlights the predicted GC-rich stem-loop. **A)** The *Bordetella* phage DGR prototype, **B)** a TR matching to two VRs within the same DGR cassette, **C)** a TR containing an insertion besides the "AAC codon" of TR and does not cause frameshift, **D)** two VRs belonging to the same DGR cassette and both have long insertions, also next to the "AAC codon" of TR, and **E)** a TR-VR pair with several non-A-to-N mismatches.

IMH, IMH and inverted repeats*

Because the *Bordetella* phage and *Legionella* DGRs have been shown to require G/C, IMH, IMH* and GC-rich inverted repeats for direct homing (Figure 40) [164, 169], all DGRs of this compilation were searched for these elements in order to determine whether they are required by other DGRs. The prediction was based on either the sequential or structural features: The G/C element mostly contains G and C, and it is located at the 3' end of the VR and TR; IMH and IMH* are located downstream of the G/C element in VR and TR respectively, and they have similar but non-identical sequences. The GC-rich inverted repeats are downstream of only the VR, and they form DNA stem-loops (Figure 40) [62, 164, 293, 294].

The sequence alignments between the TR and VR (Appendix E3) were used here to search for these elements. First, IMH, IMH* and G/C were visually identified by looking for sequences at the 3' end of VR with imperfect matches downstream of a short G/C-rich sequence. The inverted repeats were detected using the nucleotide folding program Mfold [218], in which ~200 bp downstream of the VR were folded and GC-rich short stem-loops were then screened for potential inverted repeats (Table 7, Methods).

Both G/C and IMH/IMH* could be detected in almost all DGRs of this compilation (Figure 43), but only a small number of DGRs (34%, 127 of 372) could be predicted to have a stem-loop structure that corresponded to the inverted repeats (Appendix E4). Because the inverted repeats resemble a transcriptional terminator structure in terms of the formation of a stem-loop structure, it could be possible that a terminator appear downstream of the VR may be an alternative to the inverted repeats. However, through automated terminator prediction by the ARNold webserver, no evidence could be found to indicate transcriptional terminators and inverted repeats are interchangeable (data not shown). Therefore, although G/C, IMH and IMH* are more likely to be universal features and are thus expected to be always required for directional homing, the inverted repeats may either be only necessary for certain DGRs, or there is an alternative element used by the other DGRs that lack this feature.

Diversity of target genes in number and location

The presence of multiple TGs for both *Treponema* and *Nanoarchaea* DGRs [170, 296] have implied DGRs are capable of targeting multiple genes located throughout the genome. By examining the number of TGs for DGRs in this compilation, it also showed that having multiple TGs is not rare. About ~15% of DGRs (53 of 372 DGRs) were predicted to interact with more than one TG, while having was the most common (49 of the 53 DGRs). Three of the remaining four DGRs have three TGs, and the last DGR found in *Stenotrophomonas sp.* is the most extreme example containing eight TGs. However, this result was suspicious as the TGs are all short (<150 aa), close to each other (~30 bp apart), and have large regions of identical sequences (data not shown), which suggested a potential assembling error of its genomic sequence. However, the reliability of source sequences should not cause an issue for most other DGRs, because the sequence alignments of different components did not reveal other problematic DGRs other than this exception.

Following examination of "adjacent" TGs that are next to a DGR component, there are 45 DGRs in this compilation that are predicted to also interact with non-adjacent TGs, which were defined as TGs that do not locate immediately next to other DGR components but are still considered as residing in the same neighbourhood. Moreover, 10 DGRs were found to have remote TGs, which are at least 100 kb away from the cassette, although they were not proven to be functional. Interestingly, there are also a few DGRs that only have remote targets. Altogether, these observations indicate DGRs are capable of interacting with TGs *in trans*. While this has been previously demonstrated using an artificial experimental system [164], this compilation provided more examples from bioinformatically predicted putative DGRs.

Patterns of the protein domain composition in TGs

Unlike the RT, TGs from all DGRs cannot be aligned due to the lack of sequence similarity, and only subsets of TGs can be aligned near the VR if the VRs belong to the same class. In addition, TGs of the three characterised DGRs are functionally different despite all containing a VR that has a CLec fold [169, 170, 167]. TGs are thus expected to have a great diversity in terms of both sequence and function. In order to investigate this, all adjacent TG sequences in this compilation were submitted to the CDD server to

analyse their protein domain compositions, which could be used to predict their functions. This revealed 27 different protein motifs after combining overlapping motifs or those with essentially the same functions. The domain composition of the TG was used to classify the TGs. They were grouped first based on the domain corresponding to the VR, and then by the organisation of other motifs. Eight major categories were formed, named "a" to "h", which were further divided into 39 subtypes (Figure 44).

Category "a". VRs are associated to the *mtd* domain from the Mtd tail protein and this includes the *Bordetella* phage DGR. It contains 12 members, and they all belong to the VR class CLec1. It can be divided into two subtypes, one with 11 members that contain only a single *mtd* domain (a1), and the other contains a phage tail-collar fibre protein domain at the N-terminus in addition to the *mtd* domain (a2). In addition, phage-related genes have been found in the neighbourhood of almost all 12 DGRs, suggesting DGRs in this group are similar in function to the *Bordetella* phage DGR.

Category "b". This is the largest group and contains 91 members including the *Treponema* DGR [170]. All VRs belong to the CLec1 class and are associated with an FGE-sulfatase domain. "FGE" stands for "**f**ormylglycine-**g**enerating **e**nzyme", which is a sulfatase. Proteins containing this domain in eukaryotes are required for post-translational sulfatase modification [236]. In *Treponema*, this enzyme functions as an iron-dependent oxidoreductase [236]. This category further contains 14 subtypes. Most members have no additional domains but vary in the size of the region N terminal to the FGE-sulfatase domain (b1, b2). One DGR has three VRs, each containing an FGE-sulfatase domain (b3). Based on the sequence similarity between TR and VR, all three VRs appear to be diversified. The rest of the members have one or more additional domains but their small group sizes (b4-14) limited further prediction of their functions.

Category "c". This category contains three TGs belonging to VR class CLec1. Although CDD could not detect any domain, nor could a C-type lectin fold be predicted by the Phyre2 server, they remained in CLec1 class because they could be aligned well with other CLec1 VRs (data not shown). Therefore, a box labelled "CLec1" was used here to represent the domain associated to the VR.

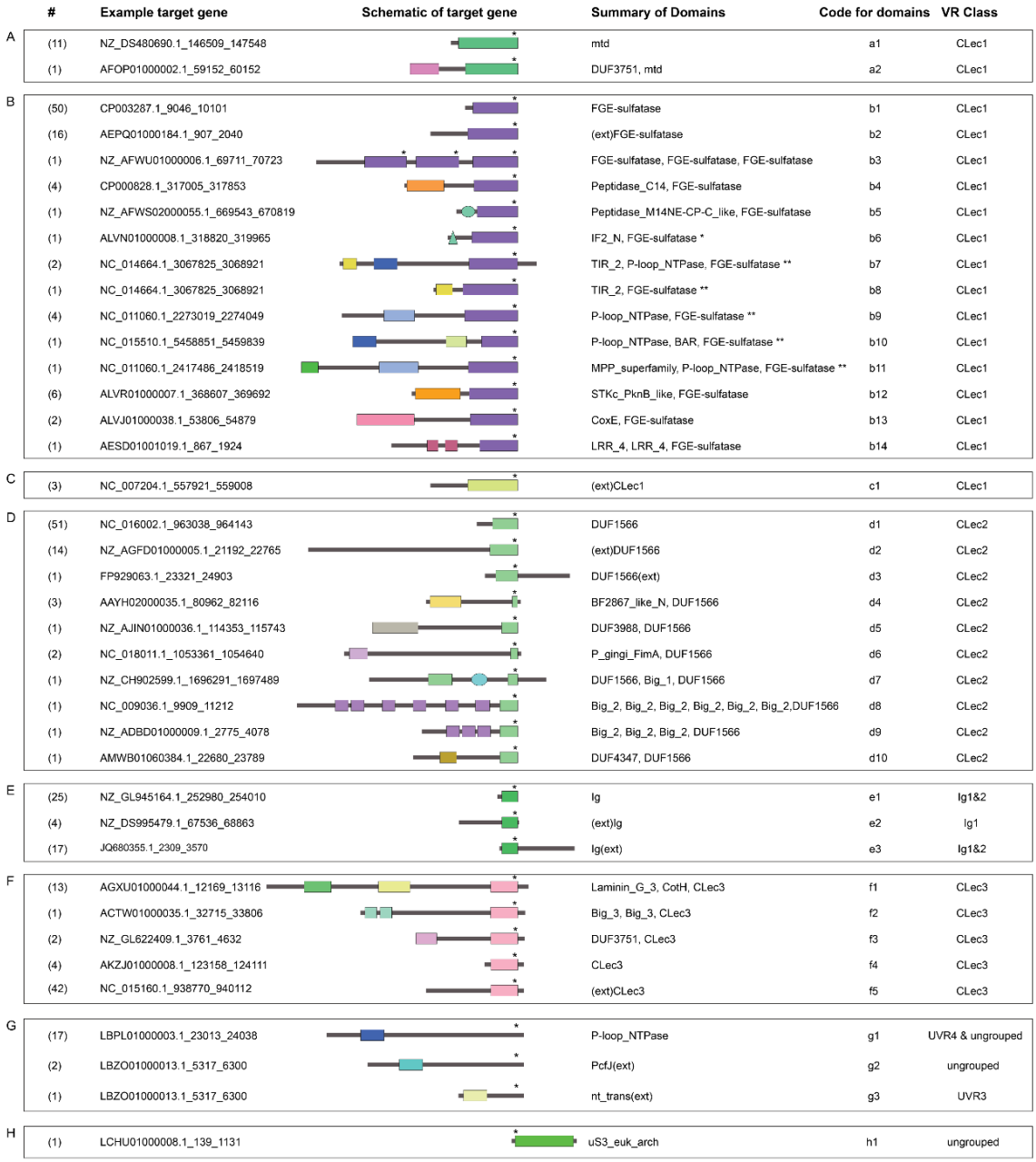


Figure 44. Domain composition of TGs.

TGs are categorised based on the domain composition, which was detected by CDD. Boxes A to H correspond to categories "a" to "h" respectively. The total number of each motif pattern is given in parentheses, followed by an example DGR. On the right, the associated VR-based class is given.

Category "d". Including the *Legionella* DGR, all DGRs in this category have a CLec2 VR that is associated with a DUF1566 domain ("DUF" stands for "**d**omain of **u**nknown **f**unction"). This domain is similar in sequence to the "Fib_succ_major" domain, which is usually related to a lipoprotein signal sequence and is thus consistent with the *Legionella* DGR [236, 169]. Like category "b", the majority of these do not contain additional domains (d1-3), however, 10 members do (d4-10). Some groups have a few bacterial Ig-like domains (e.g. Big_2, Big_3) located internal to the TG, but unlike the Ig classes of VRs, none of these domains showed evidence to be alternative VRs.

Category "e". This category resembles category "c", in which no domains nor Ig fold could be detected for the VR although a few had weak matches to domains belonging to the Ig superfamily. Therefore, they were grouped together and labelled as "Ig" to indicate their sequences could be aligned to the Ig domain.

Category "f". All members in this category belong to the VR class CLec3. Most subgroups are either small (f2, f3) or do not contain any other known domains (f4, f5). Among all subgroups, f1 is relatively large and contains 11 members that have high sequence similarity. Although matches to either the "Laminin_G_3" (Concanavalin A-like lectin/glucanases superfamily) or "CotH" (spore coat protein H) domain were not obtained for all members, both domains are drawn because of the sequence similarity. Both domains are associated with structural or binding functions, and this category is another example consistent with those experimentally characterised DGRs.

Categories "g". All members from this category are from the CPR set and they all have VRs from a UVR class or are unclassified. Neither CDD nor Phyre2 showed matches to any known protein motifs around the VR. Instead, a few matches were detected at the N-terminus. Therefore, these TGs were grouped together as a special group, in which the VR associates with no known protein motifs.

Category "h". There is only one member in this category. Different from other TGs, this TG contains the VR at the N-terminus. The entire protein contained a eukaryotic/archaeal type S3 ribosomal domain "uS3_euk_arch", giving the possibility that a DGR could be interacting with protein synthesis or related machinery.

So far, categories described above mainly included TGs from the core set (except for categories "g" and "h"), while most TGs from the CPR set matched to no known motifs. Therefore, PSI-BLAST searches were performed in order to find more distantly related

proteins to these TGs. However, this only revealed UVR1 and UVR4 to be related to DUF1127 and Midasin respectively, and the rest of UVRs and unclassified VRs remained unclear in their domain composition. The Midasin domain is an AAA-ATPase motif belonging to the AAA_16 family, which has also been seen in category "b". However, because neither of UVR1 and UVR4 could be predicted to have a CLec or Ig fold for the VR, the similarity in motif organisation could not reassign them to any of the current VR based classes.

Consistent with the VR classes that showed correlations to the RT-based tree, different TG categories correlate to the tree as shown in Appendix F2. The best examples are category "e" that contains TGs from the two Ig classes that correlate to Lineages 1 and 2, category "f" that correlates to Lineage 3, and category "d" that correlates to a large single clade (Appendix F2). Also, categories "a", "b" and "c" correlate with either large or highly supported clades elsewhere in the tree (Appendix F2). In addition, correlations were also often observed at the subgroup level (e.g. "a1", "a2") (data not shown). Although categories "g" and "h" were not correlated to any specific clade, this agrees with the fact that they belong to the CPR set, which was expected to be very diverse and had very little resolution. Because of a larger variety in domain composition compared to the number of VR based classes, the TG as a whole appears to evolve faster than the VR portion and thus could gain additional domains and extra functions.

Accessory genes

Before I started this project, four accessory genes were already known: AVD, HRDC, MSL and CH1. While the former two were previously reported in many publications, the latter two were identified in the old dataset by Michael Abebe. Later when I collected the CPR set, I used sequence profiles of the four known accessory genes to detect accessory genes of DGRs from the CPR set. In addition, to identify new potential accessory genes, all ORFs next to a DGR component were subjected to a pairwise BLASTP-based comparison to group closely related proteins, which formed four **p**otential **a**ccessory **g**ene (PAG) groups.

AVD is the most common accessory gene that was found to function in nucleic acid binding and assist the mutagenic homing event [162]. They are present in 74% (275 of 372) of DGRs in this compilation and have a shared sequence similarity. Almost all AVD-

containing DGRs have only one AVD, with only one exception containing two. Unlike TGs, all AVDs identified are close to a DGR component and no remote AVDs were found. Since AVD binds both the RT and cDNA during mutagenic retrohoming in the *Bordetella* phage DGR [162, 292], it could require the AVD and RT to be physically close. A correlation between the distribution of AVD and the RT-based tree is observed, in that AVD is found in all DGRs in the tree except for Lineage 3, a subclade of Lineage 1, and a few single clades (Appendix F3). The widespread presence of AVD across the tree implies that AVD could be the oldest accessory gene incorporated into a DGR cassette.

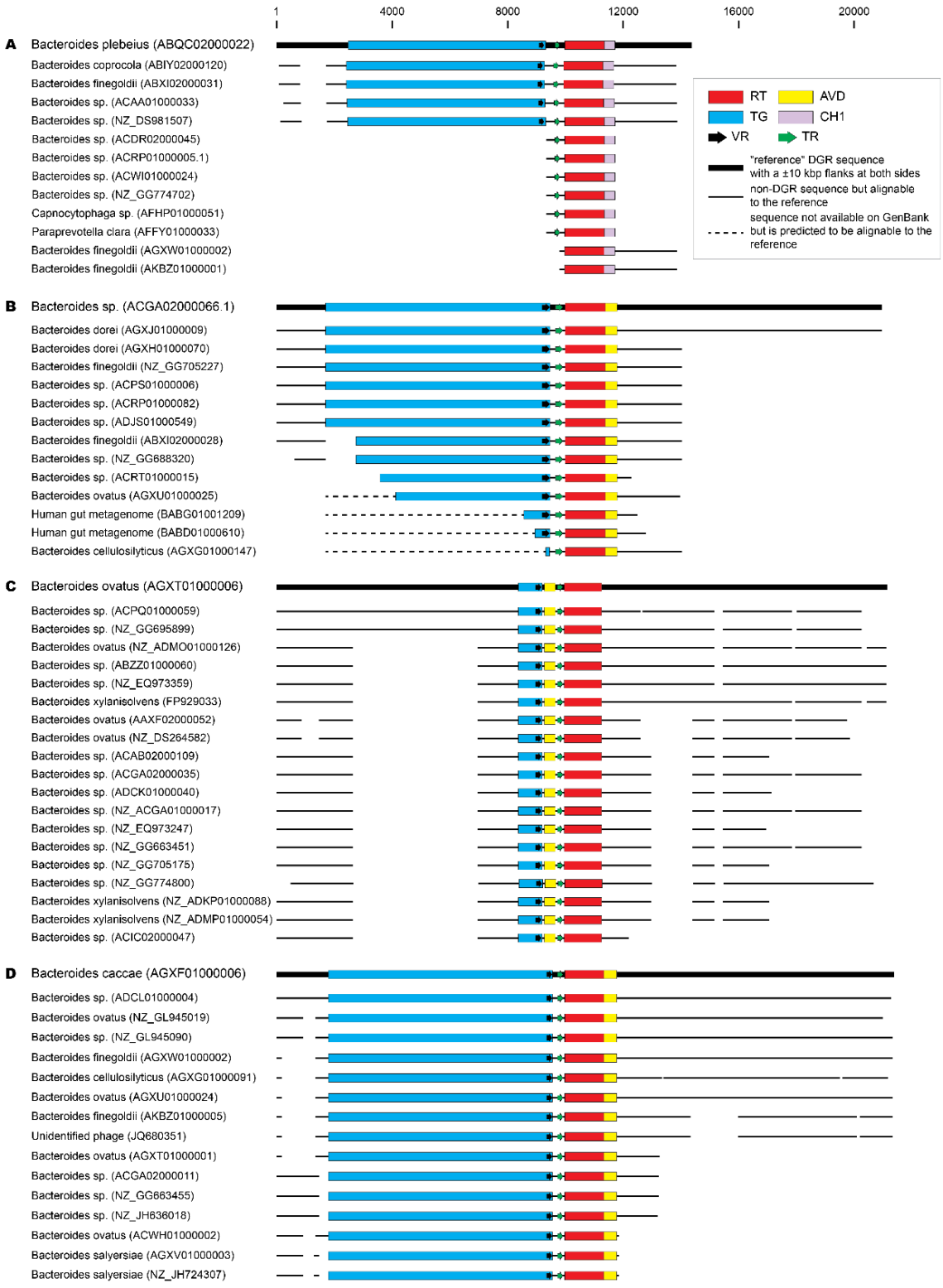
The second accessory gene is HRDC, which is found in about 4% of DGRs in this compilation (14 of 372). Like AVD, almost all HRDC-containing DGRs have only one HRDC except one example that contains two. Like AVD proteins, HRDC domains are predicted to function in nucleic acid binding [237, 236]. About half of HRDC-containing DGRs contain an AVD as well, indicating HRDC proteins may either collaborate with AVD or function independently. The distribution of HRDC correlates with the tree, in that all HRDC-containing DGRs belong to one clade, except the one that contains two HRDC's (Appendix F4). DGRs containing both HRDC and AVD branched earlier than those only containing HRDC (Appendix F3, F4), implying HRDC likely first coexisted with AVD in the same DGR cassette, but gradually completely replaced the AVD in a certain clade of DGRs.

Two new accessory genes, named MSL and CH1, were also revealed in this compilation. "MSL" stands for "**M**ut**S** **l**ike". MSL proteins are short (<300 aa) and all contain a MutS domain, which was predicted by Phyre2 and some have been annotated by GenBank as a MutS-related protein. The MutS domain is often found in mismatch binding proteins required in the DNA mismatch repair system [236]. This is consistent with both AVD and HRDC being nucleic acid binding proteins, supporting MSL proteins are coded by actual accessory genes. However, it is not yet clear whether MSL proteins function independently or together with AVD, since all MSL-containing DGRs contain an AVD as well. As stated earlier, a few RTs were also found to contain an MSL domain. Although there is no DGR that contains an MSL domain in both the RT and the accessory gene, DGRs that have an MSL domain in either component are clustered in related DGRs (Appendix F4), suggesting these DGRs could belong to a sublineage that requires the MSL domain. The MSL domain was likely acquired from the host DNA

repair system, to provide additional binding functions along with the AVD.

The last accessory gene is CH1, in which "CH" stands for "**c**onserved **h**ypothetical". They are even smaller proteins than MSL proteins (<200 aa) and are found exclusively in all 14 DGRs from Lineage 3a (Appendix F4). In contrast to MSL, no known motif could be detected in the sequence of CH1 proteins. But because CH1-containing DGRs are closely located in the tree and were not predicted to contain any other accessory gene, it is very likely that CH1 is an actual accessory gene with a yet unidentified function. This was also supported by another sequence comparison across many DGR "duplicates" from different hosts. As shown in Figure 45A, 13 DGRs that contain an RT with >=95% sequence identity were compared through BLASTN. Including the entire DGR cassette and ± 3 kb flanks at both sides, five copies are likely functional "full-length" DGRs that contain all TG, TR, RT and CH1 components, while an additional six lack the TG, and the remaining two lack both the TG and TR (Figure 45A). Although the latter eight copies that lack one or more components may be inactive remnants, CH1 is present in all 13 copies (Figure 45A), further indicating CH1 is an actual DGR component.

Finally, there are four **p**otential **a**ccessory **g**enes, named PAG 1-4. They were identified through pairwise BLASTP comparisons across all ORFs located close to a DGR component. Closely related proteins based on the E value were grouped together and assigned as PAGs if the group contained three or more members. However, they are all of a small group size (3-4 examples each) and lack enough information, such as protein motifs, to predict their functions. Except for PAG3 proteins that are limited to a single clade, the other PAGs showed no strong correlation to the RT tree (Appendix F4). These factors made it less likely that they were actual DGR components. In this case, these proteins might simply share some sequence similarity and appear to be close to a DGR cassette. For example, they could be similar genes from related phages, which have DGRs in the same location of the phage genome. Therefore, PAGs 1-4 were not considered as a solid component in the DGR cassette.



(Figure caption on the next page.)

Figure 45. Comparison of DGR cassettes with nearly identical RTs in different genomes.
The entire DGR cassette including ± 10 kb flanks were compared to a reference sequence (top) through BLASTN. Coloured boxes and arrows correspond to different DGR components, and non-DGR areas are drawn as a thick line for the reference DGR. Other DGR members in the same group are drawn below the reference and are ordered by alignable area, host organism and then GenBank accession. Non-DGR areas that are alignable to the reference are drawn as a thick line. Dashed line indicates there is no sequence available on the corresponding GenBank entry, but is predicted to be alignable to the reference. Panels A to D correspond to four individual DGR organisations.

Besides the accessory genes mentioned above, many DGRs had no accessory genes identified at all. These DGRs cluster in the RT-based tree, including a subclade of Lineage 1, most members of Lineage 3 but excluding 3a, and a small CPR clade next to Lineage 4. Accessory genes generally have nucleic acid binding activities as shown by AVD, HRDC and potentially MSL, and thus DGRs without any accessory genes may compensate for this by having an RT that can function independently without the accessory gene, requiring other proteins with a similar function from the host, or using a different mechanism to facilitate the mutagenic homing process. As stated previously, the first two possibilities could be exemplified by MSL-containing RTs or MSL accessory genes.

Meanwhile, the subclade of Lineage 3a that contain a CH1 accessory gene probably indicates the gain of CH1 during the evolution of Lineage 3 that generally have no accessory genes. Therefore, it would be possible that the initial form of DGR did not contain specifically associated accessory genes and likely had low activity. To enhance their mobile activity, most DGRs have incorporated one of the major accessory genes (AVD, HRDC, MSL, CH1) during evolution. PAGs could thus represent "intermediate" forms in DGRs that are undergoing the process of establishing their own accessory genes.

Architectures of DGR cassettes

After determining all DGR components, the organisation of each DGR cassette was depicted. To illustrate the diversity and investigate whether there is an evolutionary pattern for these architectural arrangements, DGRs were first put into major divisions based on whether an accessory gene is present and the type of accessory gene. This resulted in five major groups, named A to E (Figure 46). Next, subgroups were assigned

based on the order of RT, TR and accessory genes. The TG was neglected at this point because it was noticed that TGs could vary in both number and order, which would not contribute in a simple and reasonable way of division. This resulted in 1-3 subgroups (e.g. A1, A2) within each major group (Figure 46). Finally, both the number and location of the TG were considered to depict the specific organisation for each DGR individual, and the number of DGRs that have the same architecture was counted and shown in Figure 46.

Architecture A is the simplest group in which the core consists of the RT and TR only. It has two subgroups, A1 (TR-RT) and A2 (RT-TR), and together these DGR comprise 29% (75 of 372) of the compilation.

Architecture B has a core of RT, TR and AVD. This is the largest group and contains 68% (253 of 372) of DGRs, including the best characterised DGRs from *Bordetella* phage and *Legionella*. There are three large subgroups (B1-3) within Architecture B that differ in the order of the three core components. Interestingly, subgroup B3 has the AVD in the opposite orientation, and except one DGR, all B3 members are from the CPR set.

Architecture C has RT, TR, AVD and HRDC in the core. It contains 14 members including the characterised DGR in *Treponema*. It has one subgroup (C1) of four members representing the core of HRDC-AVD-RT-TR. The other individuals were not further divided due to the small number of representatives.

Architecture D and E contain the MSL and CH1 accessory genes, respectively. There are no subgroups in D because all core components are unique and there are not many examples. The entire group E contains only the subgroup E1, because of the same core of TR-RT-CH1.

Not only do these architectures reveal a large diversity, they also indicate there is no specific order for all DGR components. Therefore, different DGR genes can be independently expressed, rather than resembling an operon where all genes are translated together. This would allow any missing component to be obtained *in trans* from a remote genomic locus or located in the opposite orientation. Still, the general trend is that the entire DGR cassette evolves as a single unit, as can be seen from the good correlation between different architectures and the phylogenetic tree of RTs, in which each major clade mainly corresponds to only one specific architecture (Appendix F5).

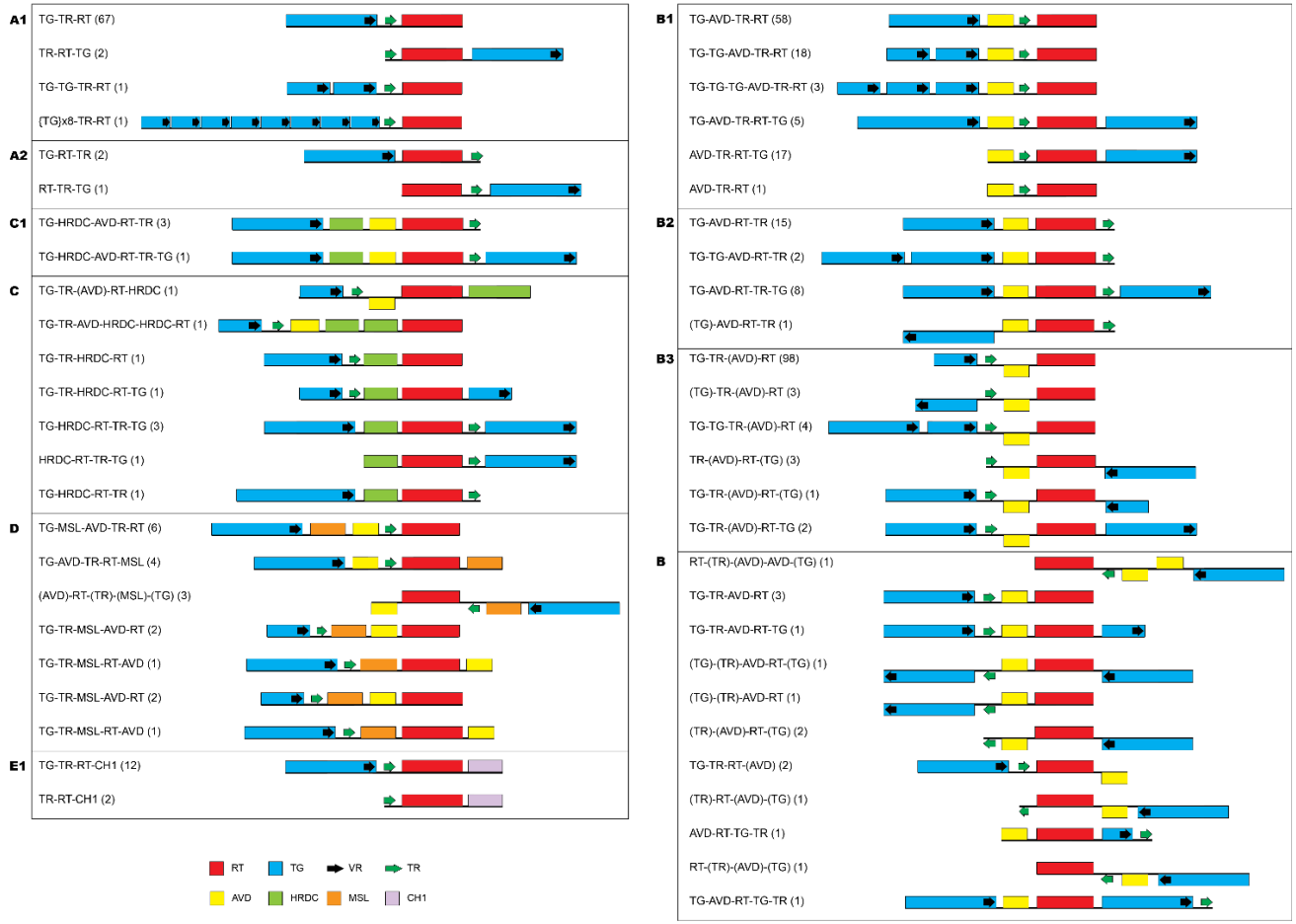


Figure 46. Architectural groupings of DGR cassettes.

Each DGR component is depicted as coloured blocks or arrows. DGR components being placed above or below the black line in the middle indicate they are in the same or opposite orientation to the RT respectively. To save the space, components are not drawn to scale, and the entire group B is arranged on the right out of the alphabetical order because of its large variety. The noted architecture is given beside each schematic and components in parentheses indicate the reverse orientation. The number of DGRs in each group is given after the noted architecture.

To further investigate whether the DGR cassette remains intact while being transferred across hosts, multiple DGR "duplicates" that have an RT of >= 95% sequence identity were compared to each other using BLASTN. Four independent sets of DGRs were subjected in this comparison and each set contained more than 10 DGR duplicates. The comparison included the entire DGR cassette and ± 10 bk flanks at both sides. As shown in Figure 45, the DGR cassette itself is intact in most hosts, but their flanking areas are often unalignable. Therefore, DGRs are usually inherited as a unit, but incomplete cassettes may still be active by function *in trans*.

Phage association

The well characterised DGRs from *Bordetella* phage, *Legionella* and *Treponema* have demonstrated that DGRs can benefit either a phage or bacterial host. Of this compilation, whether or not a DGR is related to a phage was thus investigated. The first approach was directly based on the GenBank annotation of whether the source of the sequence belonged to a phage genome. However, only 87 DGRs in this compilation were clearly annotated, most of which (69%, 60 of 87) are from a chromosome source, and smaller proportions are from free phages (24%, 21 of 87) or plasmids (7%, 6 of 87). For chromosomal DGRs, it is often unclear whether they are associated to phages, due to the lack of decisive phage gene indicators and dispersed phage remnants in the genome.

To overcome this issue, this project used two approaches to judge whether a DGR is phage-associated. The first approach was based on the TG. Since the *Bordetella* phage DGR targets a Mtd protein, other DGRs that target the same protein were likely also to be associated to phage. The second approach compared all ORFs in the neighbourhood of a DGR against a database of phage genes provided by the PHASTER website [303] through BLASTP to identify phage-related genes. To reduce the chance of false prediction, only genes coding for phage structural proteins were considered, while other genes such as transcriptional factors were neglected. The results were evaluated based on whether the match was strong (E value <= 1e-20) or medium (E value <=1-5), as well as whether matches were found at both sides or only one side of the DGR. Ideally, strong matches at both sides of a DGR would indicate a clear phage association, while other cases could still be possible but with less confidence.

With more stringent criteria that only consider DGRs with matches with an E value <=1e-20 at both sides, 34 DGRs were shown to be likely associated to phage. If using a less stringent criterion that considers matches with an E value <= 1e-5 and found in at least one side of the DGR, up to 111 DGRs could be considered to be associated with phage. Altogether, this shows that DGRs do not prefer either a bacterial or a phage host, and both are common.

When mapping DGRs that were predicted to be phage-associated to the RT-based tree, it could be seen that almost all such DGRs are from the core set and they mainly cluster in Lineages 1, 2 and 3 or nearby clades, and the clade including the *Bordetella* phage DGR has the strongest evidence (Appendix F6). Overall, this implies DGRs that benefit phages are more likely to be closely related than other DGRs. Since entering the bacterial cell is an essential during phage adaptation, phage-originated DGRs are expected to resemble the characterised *Bordetella* phage DGR and target the Mtd or similar proteins that are involved in cell binding. In contrast, Lineage 4 and many other clades, including both *Treponema* and *Legionella* DGRs as well as almost the entire CPR set, showed the least evidence for phage association. Different from DGR in phages, DGRs from a bacterial host are more likely to benefit the bacterial host and target proteins involved in host cellular functions.

The evolution of DGRs

Because a DGR contains multiple components while the RT is the only common one with sequence alignability, the easiest way to construct a phylogenetic tree of DGRs would be using the RT sequence (earlier section). However, this may not represent the full evolutionary path of the DGR cassette as each component could evolve independently from the RT. Therefore, patterns observed in earlier sections (e.g. VR-based class, the type of TG) were mapped to the RT-based tree in order to investigate whether the entire DGR cassette evolved as a single unit.

As mentioned in earlier sections, many features have shown a good correlation with the RT-based tree, including the VR-based classification, the domain composition of TG, the distribution of accessory genes, and the architecture of DGR cassettes (Figure 41, Appendix F1-5). They include almost all DGR components except for the TR, which is not classified in this study but should fall into similar classes as the VR because of the

sequence similarity. Therefore, a coevolving relationship between different DGR components and RT could be concluded, which further indicates the entire DGR cassette was evolving as a single unit. Therefore, the RT-based tree should be informative enough to represent the evolutionary path of DGRs while considering all components, and closely related DGRs observed in the tree are expected to be more similar in many aspects, such as function and host organism. This agreed with the fact that DGRs that are likely benefiting phages clustered in many clades, while DGRs lacking evidence for phage association and are expected to benefit cellular functions are mainly from other clades (Appendix F6).

From the RT-based tree, it was observed that DGRs from more closely related species (at the phylum level) are located closer to each other (Appendix F7). Although not presented in figures, DGRs from the same Class, Order and Family of hosts also showed consistent correspondence to the RT-based tree. Therefore, Appendix F7 could illustrate a possible direction of how DGRs were transferred across different phyla during evolution. Almost no DGRs were observed being transferred between the CPR and core sets. This not only indicates that DGRs between the two subsets are more different from each other, but also implies there were much fewer contacts between organisms from the core and CPR organisms, showing that their DGRs might have independent histories.

Besides what is stated above, the position of VR in TG could also be inherited by DGRs during evolution, in which most DGRs with an internal VR belong to either Lineage 2 or a subclade of Lineage 1 (Appendix F8). Because the G/C and IMH elements that locate at the 3'-end of the VR are involved in directional homing, DGRs with VRs at unusual locations might thus develop a different homing mechanism, which might affect the RT sequence and result in them being more closely related as seen in Appendix F8. However, there are multiple clusters of such DGRs, and some of them are rather scattered in the tree. It was therefore possible that such unusual features were adopted by different DGRs independently, probably with various triggers from the host environment and functional requirements.

In contrast, there are aspects that are unlikely to coevolve with either the RT or other DGR components. For example, the location of a DGR on either a chromosome, a plasmid or a free phage is rather random while being mapped onto the RT-based tree (figure not shown), indicating DGRs are frequently transferred horizontally across these

carriers. Another example is whether the TR-VR pair always contain only the canonical A-to-N substitutions. As described earlier, all possible mismatches, including A-to-N and non-A-to-N substitutions as well as indels, were observed in TR-VRs from this compilation, but there was no clear pattern showing they only occur in DGRs that are clustered in the RT-based tree (figure not shown). Similarly, the number of mismatches of each type did not correlate to the tree either (figure not shown). Together, both the type and number of mismatches in the TR-VR pair were more likely caused by random mutations rather than being inheritable properties of the DGR.

Altogether, most DGR components have shown consistent distributions compared to the RT-based tree, indicating the entire DGR cassette likely evolved as a whole. Gain or loss of a component or feature might occur while a DGR is being transferred into a different host, and rearrangement could occur over time resulting in various cassette architectures. Depending on the environment, unrelated DGRs may evolve a similar feature. Even though the RT-based tree generally lacks statistical supports for most of the major clades, consistent distributions of various aspects while being mapped onto this tree have thus provided additional supports to the tree topology as well as the current division of four lineages.

Finally, this project only constructed one phylogenetic tree using the RT, which is the only common DGR component with sequence similarity. In the future, information of other aspects, such as the VR-based class and the architecture group, can be coded as "morphological" data and used together with the RT alignment (similar to morphological trees constructed in Chapter IV). Because current observations support that all DGR components evolved as a unit, it would be expected a tree combining both sequence data of the RT and morphological data of other components should be similar to the RT-based tree, but could provide more insight to interpret the evolutionary pathway of DGRs.

Conclusion

The project described in this chapter searched and systematically compiled a set of 372 DGRs that are unique in sequence through bioinformatic approaches. Consistent from both experimentally characterised DGRs and other candidates predicted previously, this compilation expanded our knowledge of the diversity of DGRs in the aspects of function,

gene organisation, and host organism, and can be used as a reference to direct experimental design for specific individuals. A few questions are left unanswered and could guide further experimental investigation, such as the biological function of the TG and how it is benefited by the DGR-directed sequence diversification; the function of new types of accessory genes, and how DGRs without accessory genes can be functional with alternative factors or mechanisms. The effects of noncanonical substitutions as well as their cause can be an additional topic, as they are not rare but are present in a rather large number of DGRs. Finally, DGRs from the CPR set will be of special interest to investigate in general, since not only are they distinct from other DGRs, their host organisms are also distinct from most other DGR-containing bacterial species.

Since this compilation underwent manual examinations to ensure accuracy, these verified DGRs can be used as "template" DGRs to direct future searches with increased accuracy. Programs used for these analyses can be revised and packed as a single program that is more convenient and user-friendly. Even though the automated procedures may never be as accurate as manual curations, integrating current experiences will certainly be more efficient and can reduce the manual effort required in the future.

Chapter VI. Final Conclusions

Although mostly found in eukaryotes, retroelements are also present in prokaryotes and have been identified through both experimental and computational approaches. Unlike those in eukaryotes that are often selfish DNAs and associated with mobility activities, many prokaryotic RTs have been predicted to have unique functions and there is some evidence that they are even beneficial to the host cells.

In this dissertation, I performed four projects focusing on data mining and bioinformatic analyses of retroelements in bacteria, especially group II introns and DGRs. Unlike experimental approaches, bioinformatics-driven approaches can massively expand the scale of investigations at the genomic or metagenomic levels, infer evolutionary histories and even reproduce the progress of evolution through computational simulation. This dissertation took advantage of the enormous sequence databases and various existing bioinformatic tools to carry out several projects that gathered fundamental information about bacterial retroelements.

Overall, the basic framework used for data mining was of similar design for the different projects, with modifications made to fit different characteristics of each type of element. For example, all projects started by identifying RTs, using a protein profile for the type (e.g. profiles created from alignments of general RTs, group II-related RTs, or DGR-related RTs). Additional prediction procedures were applied individually to specific projects to further refine the results. In the case of group II introns, the RNA secondary structures were predicted through RNA-based motif searches in the flanks of each group II RT candidate. Meanwhile, RNA-based motif searches also played an essential role in finding non-standard intron organisations as described in Chapter III. In the case of DGRs, additional components other than the RT were identified in a hierarchical order of TR-VR, TG and the accessory gene(s), since the A-to-N mutagenesis in the TR-VR pair is the most characteristic and can be predicted easily through pairwise BLASTN. All searches involved in this dissertation were made against the GenBank's non-redundant database, but it is also possible to search other databases using the same programs with minor modifications.

The first project (Chapter II) performed a general search for bacterial retroelements. Although the majority of identified RTs were found to be group II introns, retrons and DGRs, seven groups were assigned as new types of RTs. Through examinations of the

protein motif composition, copy number and genes in the neighbourhood of the RT, most types of RTs remained unclear as to the nature of their functionality, while group II appeared to be the only type of retroelement with clear evidence of mobility. This indicates that group II introns have the highest probability of being widespread across bacterial species, and thus would have had a higher chance to be transferred into eukaryotes, consistent with the general hypothesis that group II introns were the ancestors of eukaryotic retroelements and the spliceosome.

The project described in Chapter II also demonstrates that automated searches using only standard BLAST programs were able to identify potential new RT classes. Considering that standard BLAST searches are easier to be automated compared to PSI-BLAST, future searches can adapt the demonstrated BLAST method to quickly identify sequences similar to given queries. However, whether PSI-BLAST is able to identify sequences that are more divergent from the set of currently known RTs remains uncertain, and this possibility could be tested in the future. Other possibilities for future studies concern the predicted functions of established classes. The distribution of each class in different organisms could be examined, which might provide hints about their roles in living cells in the context of the host environment.

Similar searches were specialised to collect and analyse bacterial group II introns and DGRs (Chapter III and Chapter V). The final dataset for each project included only those sequences that are unique to avoid redundancy. Many bioinformatic analyses were then performed for introns and DGRs, which provided updated and expanded information to our current understanding of these retroelements. Previously, our database of group II introns mainly contained intron prototypes that are unique in sequence relative to each other. In Chapter III, the database was expanded by recording multiple copies of each intron prototype, non-standard intron compound organisations, and examples of intron graveyards. Some of such organisations were observed multiple times in one or even different genomes, and thus could be possible candidates to be further experimental characterisation.

The updated intron collection is also expected to better represent the entire intron population in nature. By combining knowledge of copies in different genomic sites, another future topic could be to investigate how group II introns became distributed within the same host cell, the same host species, or across different host species. In addition, although not described in this dissertation, some data collected in this project

gave rise to a hypothesis for a productive experimental project focusing on the intron-exon recognition in class A introns, which was performed by Ashley Jarding in our lab. It is expected that the assembled dataset will generate additional hypotheses for experiments in the future.

In the technical aspect, the project described in Chapter III will contribute to refinements and improvements in the automated procedure of analysing group II introns. Importantly, the set of newly identified introns allows for greater scope and accuracy of sequence alignments of both the IEP and RNA components. Updated sequence profiles for each class can now be generated from the resulting new alignments and replace the former files used by our searching program. This is expected to not only increase accuracy during the intron identification step, but also to result in better predictions for the RNA secondary structure. In addition, some procedures that were performed manually in this dissertation can be automated or partially automated in the future by integrating the information obtained. One example of this is the detection of non-standard intron compound organisations, in which an automated approach can infer organisation of different intron elements (e.g. individual RNA domains or IEP motifs) by identifying those located close to each other and assigning an overall order resembling a known complex organisation.

The project described in Chapter III also contributed my subsequent project described in Chapter IV by providing an enlarged dataset for phylogenetic analyses of group II introns. The expanded dataset indicated that the IEPs and RNAs of group II introns coevolved as single units, although there were some disagreements between IEP- and RNA-based trees that suggested exceptions. To evaluate whether these disagreements could be actual conflicts between the IEP and RNA, I used topology tests, however, it was later found that trees of group II introns may not be suitable for standard topological analysis. Because group II introns are widespread across organisms across domains of life, they must have undergone various evolutionary events such as point mutations, indels or exchanges of sequences with other introns or genomic sequences. While these changes may introduce new structural features to introns and increase their adaptability, they also would cause a lack of sequence similarity among various introns and hinder phylogenetic analysis based on the sequence data. Generally, introns that belong to the same clade in the IEP-based tree were found to share RNA structural features, but the topologies inside each clade were less robust, especially if the clade was large.

Topologies within such clades rarely had strong support, and thus topological variations observed in independently derived trees are expected to have equal validity for the align datasets. However, topology tests do not take into account statistical support, which could be the reason that the IEP- and RNA-based trees appeared to conflict even for the simplest case of class A introns, in which a detailed inspection revealed only minor differences. An attempt to resolve this issue was to perform topology tests using collapsed trees (e.g. nodes with a bootstrap < 75% were collapsed), but unexpectedly, even more conflicts were uncovered. Therefore, the topology tests were considered less suitable for evaluating coevolution of IEPs and RNAs of group II introns that are more diverse in sequence and less robust in topology. Manual comparisons of tree topologies were performed in this project, but it would be beneficial to seek alternative methods to evaluate phylogenetic trees suffering from such issues.

A novel approach not performed previously for group II introns was the construction of morphology-based trees. This approach successfully provided additional evidence for the currently inferred relationship between major classes that otherwise lacked statistical support. It might be argued that the steps of identifying and categorising morphological features involved my own judgment and might be subjective. However, the most recent RNA structural alignment for each class (refined along with Chapter III) was used during the grouping procedure to minimise the effects of bias. This time, each class was evaluated as one unit because these steps were time consuming and required a large amount of manual effort. An ideal situation would be using automatization to define and score different structural features, which would allow all introns to be included in the analysis and thus make the results directly comparable to sequence-based trees. However, this involves the correct differentiation among simple stem-loops in different lengths, stem-loops with internal loops or bulges, and/or the combination of many basic structural motifs. To correctly program such differentiation is unlikely to be achieved in a short term. Instead, as a more practical approach, the current morphological tree can be further refined by adding more examples selected from major clades of each class, and hopefully more insight can be gained to interpret the class-wide relationships, including the relationship between the two CL classes.

In summary, by using various bioinformatic tools, not only has my work successfully demonstrated the utility of data mining of specific bacterial retroelements, but also

provided insights to our understanding of retroelement distribution, diversity in structures and functions, as well as their evolutionary histories. At present, bioinformatic tools have been incorporated in almost every field of biological research. While various tools are available for various tasks, each tool has its own advantages and disadvantages, and manual efforts still play an essential role to obtain high accuracy. Although absolute perfection may never be achieved in this regard, it is expected that additional bioinformatic tools in the future coupled with the rapidly growing machine learning methods will make predictions more accurate while requiring only a minimal amount of human-directed supervision.

References

1. Crick F. On Protein Synthesis. *In F.K. Sanders. Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules. Cambridge University Press.* 1958.
2. Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature.* 1970;**226**(5252):1209-11.
3. Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature.* 1970;**226**(5252):1211-3.
4. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Gene.* 2011;**7**(12):e1002384.
5. Platt RN 2nd, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res.* 2018;**26**(1-2):25-43.
6. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol.* 2015;**86**:90-109.
7. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007;**8**(12):973-82.
8. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA.* 2017;**8**:19.
9. Boeke JD, Corces VG. Transcription and reverse transcription of retrotransposons. *Annu Rev Microbiol.* 1989;**43**:403-34.
10. Finnegan DJ. Eukaryotic transposable elements and genome evolution. *Trends Genet.* 1989;**5**(4):103-7.
11. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. *Proc Natl Acad Sci U S A.* 2006;**103**(25):9590-4.
12. Temin HM. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol Biol Evol.* 1985;**2**(6):455-68.
13. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM. Transposons in place of telomeric repeats at a Drosophila telomere. *Cell.* 1993;**75**(6):1083-93.
14. Gladyshev EA, Arkhipova IR. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A.* 2011;**108**(51):20311-6.
15. Lampson BC, Inouye M, Inouye S. Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res.* 2005;**110**(1-4):491-9.
16. Lim D, Maas WK. Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA-RNA compound in E. coli B. *Cell.* 1989;**56**(5):891-904.
17. Lampson BC, Inouye M, Inouye S. Reverse transcriptase with concomitant ribonuclease H activity in the cell-free synthesis of branched RNA-linked msDNA of Myxococcus xanthus. *Cell.* 1989;**56**(4):701-7.
18. Lambowitz AM, Zimmerly S. Mobile group II introns. *Annu Rev Genet.* 2004;**38**:1-35.
19. Medhekar B, Miller JF. Diversity-generating retroelements. *Curr Opin Microbiol.* 2007;**10**(4):388-95.
20. Simon DM, Zimmerly S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* 2008;**36**(22):7219-29.
21. Kojima KK, Kanehisa M. Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol Biol Evol.* 2008;**25**(7):1395-404.
22. Zimmerly S, Wu L. An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr.* 2015;**3**(2):MDNA3-0058-2014.

23. Nisole S, Saïb A. Early steps of retrovirus replicative cycle. *Retrovirology.* 2004;**1**:9.
24. Zhang X, Ma X, Jing S, Zhang H, Zhang Y. Non-coding RNAs and retroviruses. *Retrovirology.* 2018;**15**(1):20.
25. Finnegan DJ. Retrotransposons. *Curr Biol.* 2012;**22**(11):R432-7.
26. Galligan JT, Kennell JC. Retroplasmids: Linear and Circular Plasmids that Replicate via Reverse Transcription. *In: Meinhardt F., Klassen R. (eds) Microbial Linear Plasmids. Microbiology Monographs, vol 7. Springer, Berlin, Heidelberg.* 2007.
27. Thompson PJ, Macfarlan TS, Lorincz MC. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Mol Cell.* 2016;**62**(5):766-76.
28. Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B, Vaury C. Life cycle of an endogenous retrovirus, ZAM, in Drosophila melanogaster. *J Virol.* 2000;**74**(22):10658-69.
29. Kim A, Terzian C, Santamaria P, Pélisson A, Purd'homme N, Bucheton A. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. *Proc Natl Acad Sci U S A.* 1994;**91**(4):1285-9.
30. Nefedova L, Kim A. Mechanisms of LTR-Retroelement Transposition: Lessons from Drosophila melanogaster. *Viruses.* 2017;**9**(4). pii: E81.
31. Arkhipova IR. Distribution and phylogeny of Penelope-like elements in eukaryotes. *Syst Biol.* 2006;**55**(6):875-85.
32. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* 1990;**9**(10):3353-62.
33. Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol.* 1999;**16**(6):793-805.
34. Kapitonov VV, Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene.* 2009;**448**(2):207-13.
35. Singer MF. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell.* 1982;**28**(3):433-4.
36. Jurka J. Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol.* 1998;**8**(3):333-7.
37. Petrov DA, Hartl DL. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol Biol Evol.* 1998;**15**(3):293-302.
38. Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* 2002;**21**(21):5899-910.
39. Schumann GG, Gogvadze EV, Osanai-Futahashi M, Kuroki A, Münk C, Fujiwara H, Ivics Z, Buzdin AA. Unique functions of repetitive transcriptomes. *Int Rev Cell Mol Biol.* 2010;**285**:115-88.
40. Ostertag EM, Kazazian HH. Genetics: LINEs in mind. *Nature.* 2005;**435**(7044):890-1.
41. Aksoy S, Williams S, Chang S, Richards FF. SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. *Nucleic Acids Res.* 1990;**18**(4):785-92.
42. Gabriel A, Yen TJ, Schwartz DC, Smith CL, Boeke JD, Sollner-Webb B, Cleveland DW. A rapidly rearranging retrotransposon within the miniexon gene locus of Crithidia fasciculata. *Mol Cell Biol.* 1990;**10**(2):615-24.
43. Burke WD, Calalang CC, Eickbush TH. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol.* 1987;**7**(6):2221-30.
44. Burke WD, Müller F, Eickbush TH. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res.* 1995;**23**(22):4628-34.
45. Xiong Y, Eickbush TH. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol Cell Biol.* 1988;**8**(1):114-23.

46. Kramerov DA, Vassetzky NS. Short retroposons in eukaryotic genomes. *Int Rev Cytol.* 2005;**247**:165-221.
47. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet.* 2003;**35**(1):41-8.
48. Kajikawa M, Okada N. LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell.* 2002;**111**(3):433-44.
49. Rowold DJ, Herrera RJ. Alu elements and the human genome. *Genetica.* 2000;**108**(1):57-72.
50. Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res.* 2008;**16**(1):203-15.
51. Hancks DC, Kazazian HH Jr. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev.* 2002;**22**(3):191-203.
52. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet.* 2002;**3**(5):370-9.
53. Deininger P. Alu elements: know the SINEs. *Genome Biol.* 2011;**12**(12):236.
54. Deininger PL, Batzer MA. Alu repeats and human disease. *Mol Genet Metab.* 1999;**67**(3):183-93.
55. Machado PM, Brandão RD, Cavaco BM, Eugénio J, Bento S, Nave M, Rodrigues P, Fernandes A, Vaz F. Screening for a BRCA2 rearrangement in high-risk breast/ovarian cancer families: evidence for a founder effect and analysis of the associated phenotypes. *J Clin Oncol.* 2007;**25**(15):2027-34.
56. Belancio VP, Hedges DJ, Deininger P. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res.* 2008;**18**(3):343-58.
57. Cervera A, De la Peña M. Eukaryotic penelope-like retroelements encode hammerhead ribozyme motifs. *Mol Biol Evol.* 2014;**31**(11):2941-7.
58. Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. Retroelements containing introns in diverse invertebrate taxa. *Nat Genet.* 2003;**33**(2):123-4.
59. Evgen'ev MB, Arkhipova IR. Penelope-like elements–a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res.* 2005;**110**(1-4):510-21.
60. Lyozin GT, Makarova KS, Velikodvorskaja VV, Zelentsova HS, Khechumian RR, Kidwell MG, Koonin EV, Evgen'ev MB. The structure and evolution of Penelope in the virilis species group of Drosophila: an ancient lineage of retroelements. *J Mol Evol.* 2001;**52**(5):445-56.
61. Pyatkov KI, Arkhipova IR, Malkova NV, Finnegan DJ, Evgen'ev MB. Reverse transcriptase and endonuclease activities encoded by Penelope-like retroelements. *Proc Natl Acad Sci U S A.* 2004;**101**(41):14719-24.
62. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature.* 2004;**431**(7007):476-81.
63. Chang GS, Hong Y, Ko KD, Bhardwaj G, Holmes EC, Patterson RL, van Rossum DB. Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity. *Proc Natl Acad Sci U S A.* 2008;**105**(36):13474-9.
64. Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res.* 2007;**134**(1-2):221-34.
65. Gladyshev EA, Arkhipova IR. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A.* 2007;**104**(22):9352-7.
66. Cappello J, Handelsman K, Lodish HF. Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell.* 1985;**43**(1):105-15.

67. Goodwin TJ, Poulter RT. The DIRS1 group of retrotransposons. *Mol Biol Evol.* 2001;**18**(11):2067-82.
68. Goodwin TJ, Poulter RT. A new group of tyrosine recombinase-encoding retrotransposons. *Mol Biol Evol.* 2004;**21**(4):746-59.
69. Toh H, Hayashida H, Miyata T. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. *Nature.* 1983;**305**(5937):827-9.
70. Menéndez-Arias L, Sebastián-Martín A, Álvarez M. Viral reverse transcriptases. *Virus Res.* 2017;**234**:153-176.
71. Summers J, Mason WS. Replication of the genome of a hepatitis B–like virus by reverse transcription of an RNA intermediate. *Cell.* 1982;**29**(2):403-15.
72. Hohn T, Rothnie H. Plant pararetroviruses: replication and expression. *Curr Opin Virol.* 2013;**3**(6):621-8.
73. Nassal M. Hepatitis B viruses: reverse transcription a different way. *Virus Res.* 2008;**134**(1-2):235-49.
74. Clark DN, Hu J. Hepatitis B virus reverse transcriptase - Target of current antiviral therapy and future drug development. *Antiviral Res.* 2015;**123**:132-7.
75. O'Connor CM, Lai CK, Collins K. Two purified domains of telomerase reverse transcriptase reconstitute sequence-specific interactions with RNA. *J Biol Chem.* 2005;**280**(17):17533-9.
76. Robart AR, Collins K. Human telomerase domain interactions capture DNA for TEN domain-dependent processive elongation. *Mol Cell.* 2011;**42**(3):308-18.
77. Rouda S, Skordalakes E. Structure of the RNA-binding domain of telomerase: implications for RNA recognition and binding. *Structure.* 2007;**15**(11):1403-12.
78. Dey A, Chakrabarti K. Current Perspectives of Telomerase Structure and Function in Eukaryotes with Emerging Views on Telomerase in Human Parasites. *Int J Mol Sci.* 2018;**19**(2):333.
79. Belfort M, Curcio MJ, Lue NF. Telomerase and retrotransposons: reverse transcriptases that shaped genomes. *Proc Natl Acad Sci U S A.* 2011;**108**(51):20304-10.
80. Hoffman H, Rice C, Skordalakes E. Structural Analysis Reveals the Deleterious Effects of Telomerase Mutations in Bone Marrow Failure Syndromes. *J Biol Chem.* 2017;**292**(11):4593-4601.
81. Boeke JD. The unusual phylogenetic distribution of retrotransposons: a hypothesis. *Genome Res.* 2003;**13**(9):1975-83.
82. Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science.* 1997;**276**(5312):561-7.
83. Eickbush TH. Telomerase and retrotransposons: which came first?. *Science.* 1997;**277**(5328):911-2.
84. Weinrich SL, Pruzan R, Ma L, Ouellette M, Tesmer VM, Holt SE, Bodnar AG, Lichtsteiner S, Kim NW, Trager JB, Taylor RD, Carlos R, Andrews WH, Wright WE, Shay JW, Harley CB, Morin GB. Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT. *Nat Genet.* 1997;**17**(4):498-502.
85. Shampay J, Blackburn EH. Generation of telomere-length heterogeneity in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A.* 1988;**85**(2):534-8.
86. Poole JC, Andrews LG, Tollefsbol TO. Activity, function, and gene regulation of the catalytic subunit of telomerase (hTERT). *Gene.* 2001;**269**(1-2):1-12.
87. de Lange T. How telomeres solve the end-protection problem. *Science.* 2009;**326**(5955):948-52.
88. Greider CW, Blackburn EH. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell.* 1987;**51**(6):887-98.

89. Greider CW, Blackburn EH. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. *Nature.* 1989;**337**(6205):331-7.

90. Yu GL, Bradley JD, Attardi LD, Blackburn EH. In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs. *Nature.* 1990;**344**(6262):126-32.

91. Dokal I. Dyskeratosis congenita. *Hematology Am Soc Hematol Educ Program.* 2011;**2011**:480-6.

92. Yushenova IA, Arkhipova IR. Biochemical properties of bacterial reverse transcriptase-related (rvt) gene products: multimerization, protein priming, and nucleotide preference. *Curr Genet.* 2018;[Epub ahead of print].

93. Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol.* 2011;**3**(8):a003616.

94. Lehmann K, Schmidt U. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol.* 2003;**38**(3):249-303.

95. Lang BF, Ahne F, Bonen L. The mitochondrial genome of the fission yeast Schizosaccharomyces pombe. The cytochrome b gene has an intron closely related to the first two introns in the Saccharomyces cerevisiae cox1 gene. *J Mol Biol.* 1985;**184**(3):353-66.

96. Zimmer M, Welser F, Oraler G, Wolf K. Distribution of mitochondrial introns in the species Schizosaccharomyces pombe and the origin of the group II intron in the gene encoding apocytochrome b. *Curr Genet.* 1987;**12**(5):329-36.

97. Dai L, Zimmerly S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.* 2002;**30**(5):1091-102.

98. Dai L, Zimmerly S. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA.* 2003;**9**(1):14-9.

99. Toro N. Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environ Microbiol.* 2003;**5**(3):143-51.

100. Toro N, Jiménez-Zurdo JI, García-Rodríguez FM. Bacterial group II introns: not just splicing. *FEMS Microbiol Rev.* 2007;**31**(3):342-58.

101. Vallès Y, Halanych KM, Boore JL. Group II introns break new boundaries: presence in a bilaterian's genome. *PLoS One.* 2008;**3**(1):e1488.

102. Michel F, Dujon B. Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members. *EMBO J.* 1983;**2**(1):33-8.

103. Michel F, Ferat JL. Structure and activities of group II introns. *Annu Rev Biochem.* 1995;**64**:435-61.

104. Qin PZ, Pyle AM. The architectural organization and mechanistic function of group II intron structural elements. *Curr Opin Struct Biol.* 1998;**8**(3):301-8.

105. Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev.* 1997;**11**(21):2910-24.

106. Zimmerly S, Hausner G, Wu Xc. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.* 2001;**29**(5):1238-50.

107. San Filippo J, Lambowitz AM. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol.* 2002;**324**(5):933-51.

108. Zimmerly S, Guo H, Perlman PS, Lambowitz AM. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell.* 1995;**82**(4):545-54.

109. Robart AR, Zimmerly S. Group II intron retroelements: function and diversity. *Cytogenet Genome Res.* 2005;**110**(1-4):589-97.

110. Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM. A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol.* 2004;**340**(2):211-31.

111. Mohr G, Perlman PS, Lambowitz AM. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.* 1993;**21**(22):4991-7.

112. Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry.* 1999;**38**(28):9069-83.

113. Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell.* 1999;**4**(2):239-50.

114. Matsuura M, Noah JW, Lambowitz AM. Mechanism of maturase-promoted group II intron splicing. *EMBO J.* 2001;**20**(24):7259-70.

115. Singh RN, Saldanha RJ, D'Souza LM, Lambowitz AM. Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. *J Mol Biol.* 2002;**318**(2):287-303.

116. Lambowitz AM, Perlman PS. Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing. *Trends Biochem Sci.* 1990;**15**(11):440-4.

117. Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. A self-splicing RNA excises an intron lariat. *Cell.* 1986;**44**(2):213-23.

118. van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA. Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell.* 1986;**44**(2):225-34.

119. Podar M, Perlman PS, Padgett RA. The two steps of group II intron self-splicing are mechanistically distinguishable. *RNA.* 1998;**4**(8):890-900.

120. Daniels DL, Michels WJ Jr, Pyle AM. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. *J Mol Biol.* 1996;**256**(1):31-49.

121. Fedorova O, Zingler N. Group II introns: structure, folding and splicing mechanism. *Biol Chem.* 2007;**388**(7):665-78.

122. Marcia M, Pyle AM. Visualizing group II intron catalysis through the stages of splicing. *Cell.* 2012;**151**(3):497-507.

123. Marcia M, Somarowthu S, Pyle AM. Now on display: a gallery of group II intron structures at different stages of catalysis. *Mob DNA.* 2013;**4**(1):14.

124. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell.* 1993;**72**(4):595-605.

125. Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell.* 1995;**83**(4):529-38.

126. Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell.* 1998;**94**(4):451-62.

127. Guo H, Zimmerly S, Perlman PS, Lambowitz AM. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J.* 1997;**16**(22):6835-48.

128. Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev.* 2005;**19**(20):2477-87.

129. Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J, Lambowitz AM, Belfort M. Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol Microbiol.* 2005;**56**(2):509-24.

130. Yao J, Truong DM, Lambowitz AM. Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. *PLoS Genet.* 2013;**9**(4):e1003469.

131. Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji TO, Morimura K, Ikeda H, Hattori M, Beppu T. Genome sequence of Symbiobacterium thermophilum, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res.* 2004;**32**(16):4937-44.

132. Zhong J, Lambowitz AM. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J.* 2003;**22**(17):4555-65.

133. Martínez-Abarca F, Barrientos-Durán A, Fernández-López M, Toro N. The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res.* 2004;**32**(9):2880-8.

134. Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol.* 2002;**46**(5):1259-72.

135. Martínez-Abarca F, García-Rodríguez FM, Toro N. Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Mol Microbiol.* 2000;**35**(6):1405-12.

136. Toro N, Martínez-Abarca F. Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. *PLoS One.* 2013;**8**(1):e55102.

137. Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc Natl Acad Sci U S A.* 2001;**98**(23):13207-12.

138. Martínez-Abarca F, Toro N. RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucleic Acids Res.* 2000;**28**(21):4397-402.

139. Mueller MW, Allmaier M, Eskes R, Schweyen RJ. Transposition of group II intron aI1 in yeast and invasion of mitochondrial genes at new locations. *Nature.* 1993;**366**(6451):174-6.

140. Eskes R, Liu L, Ma H, Chao MY, Dickson L, Lambowitz AM, Perlman PS. Multiple homing pathways used by yeast mitochondrial group II introns. *Mol Cell Biol.* 2000;**20**(22):8432-46.

141. Sorek R, Kunin V, Hugenholtz P. CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol.* 2008;**6**(3):181-6.

142. Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol Cell.* 2014;**54**(2):234-44.

143. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science.* 2007;**315**(5819):1709-12.

144. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet.* 2010;**11**(3):181-90.

145. Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol.* 2014;**21**(6):528-34.

146. Silas S, Makarova KS, Shmakov S, Páez-Espino D, Mohr G, Liu Y, Davison M, Roux S, Krishnamurthy SR, Fu BXH, Hansen LL, Wang D, Sullivan MB, Millard A, Clokie MR, Bhaya D, Lambowitz AM, Kyrpides NC, Koonin EV, Fire AZ. On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *MBio.* 2017;**8**(4):e00897-17.

147. Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One.* 2014;**9**(11):e114083.

148. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* 2014;**42**(10):6091-105.

149. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol.* 2014;**12**(7):479-92.

150. Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, Lambowitz AM, Fire AZ. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science.* 2016;**351**(6276):aad4234.

151. Koonin EV, Makarova KS. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. *Genome Biol Evol.* 2017;**9**(10):2812-2825.

152. Lampson B, Inouye M, Inouye S. The msDNAs of bacteria. *Prog Nucleic Acid Res Mol Biol.* 2001;**67**:65-91.

153. Inouye S, Inouye M. The retron: a bacterial retroelement required for the synthesis of msDNA. *Curr Opin Genet Dev.* 1993;**3**(5):713-8.

154. Inouye M, Inouye S. msDNA and bacterial reverse transcriptase. *Annu Rev Microbiol.* 1991;**45**:163-86.

155. Yee T, Furuichi T, Inouye S, Inouye M. Multicopy single-stranded DNA isolated from a gram-negative bacterium, Myxococcus xanthus. *Cell.* 1984;**38**(1):203-9.

156. Darmon E, Leach DR. Bacterial genome instability. *Microbiol Mol Biol Rev.* 2014;**78**(1):1-39.

157. Inouye M, Ke H, Yashio A, Yamanaka K, Nariya H, Shimamoto T, Inouye S. Complex formation between a putative 66-residue thumb domain of bacterial reverse transcriptase RT-Ec86 and the primer recognition RNA. *J Biol Chem.* 2004;**279**(49):50735-42.

158. Inouye S, Hsu MY, Xu A, Inouye M. Highly specific recognition of primer RNA structures for 2'-OH priming reaction by bacterial reverse transcriptases. *J Biol Chem.* 1999;**274**(44):31236-44.

159. Elfenbein JR, Knodler LA, Nakayasu ES, Ansong C, Brewer HM, Bogomolnaya L, Adams LG, McClelland M, Adkins JN, Andrews-Polymenis HL. Multicopy Single-Stranded DNA Directs Intestinal Colonization of Enteric Pathogens. *PLoS Genet.* 2015;**11**(9):e1005472.

160. Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics.* 2012;**13**:430.

161. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL, Du A, Guo H, Ghosh P, Miller JF, Zimmerly S. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res.*2018;**46**(1):11-24.

162. Alayyoubi M, Guo H, Dey S, Golnazarian T, Brooks GA, Rong A, Miller JF, Ghosh P. Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase. *Structure.* 2013;**21**(2):266-76.

163. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J, Miller JF. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science.* 2002;**295**(5562):2091-4.

164. Guo H, Tse LV, Barbalat R, Sivaamnuaiphorn S, Xu M, Doulatov S, Miller JF. Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Mol Cell.* 2008;**31**(6):813-23.

165. Melvin JA, Scheller EV, Miller JF, Cotter PA. Bordetella pertussis pathogenesis: current and future challenges. *Nat Rev Microbiol.* 2014;**12**(4):274-88.

166. Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, Baker S, Davis P, Simmonds M, Churcher C, Mungall K, Quail MA, Preston A, Harvill ET, Maskell DJ, Eiserling FA, Parkhill J, Miller JF. Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol.* 2004;**186**(5):1503-17.

167. McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, Marti-Renom MA, Doulatov S, Narayanan E, Sali A, Miller JF, Ghosh P. The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol.* 2005;**12**(10):886-92.

168. Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. Selective ligand recognition by a diversity-generating retroelement variable protein. *PLoS Biol.* 2008;**6**(6):e131.

169. Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, Czornyj E, Liu M, Dey S, Ghosh P, Miller JF. Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. *Proc Natl Acad Sci U S A.* 2013;**110**(20):8212-7.

170. Le Coq J, Ghosh P. Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. *Proc Natl Acad Sci U S A.* 2011;**108**(35):14649-53.

171. Chopin MC, Chopin A, Bidnenko E. Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol.* 2005;**8**(4):473-9.

172. Fortier LC, Bouchard JD, Moineau S. Expression and site-directed mutagenesis of the lactococcal abortive phage infection protein AbiK. *J Bacteriol.* 2005;**187**(11):3721-30.

173. Durmaz E, Klaenhammer TR. Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected Lactococcus lactis. *J Bacteriol.* 2007;**189**(4):1417-25.

174. Odegrip R, Nilsson AS, Haggård-Ljungquist E. Identification of a gene encoding a functional reverse transcriptase within a highly variable locus in the P2-like coliphages. *J Bacteriol.* 2006;**188**(4):1643-7.

175. Hill C, Miller LA, Klaenhammer TR. Nucleotide sequence and distribution of the pTR2030 resistance determinant (hsp) which aborts bacteriophage infection in lactococci. *Appl Environ Microbiol.* 1990;**56**(7):2255-8.

176. Boucher I, Emond E, Dion E, Montpetit D, Moineau S. Microbiological and molecular impacts of AbiK on the lytic cycle of Lactococcus lactis phages of the 936 and P335 species. *Microbiology.* 2000;**146**(Pt 2):445-53.

177. Tangney M, Fitzgerald GF. Effectiveness of the lactococcal abortive infection systems AbiA, AbiE, AbiF and AbiG against P335 type phages. *FEMS Microbiol Lett.* 2002;**210**(1):67-72.

178. Dinsmore PK, Klaenhammer TR. Molecular characterization of a genomic region in a Lactococcus bacteriophage that is involved in its sensitivity to the phage defense mechanism AbiA. *J Bacteriol.* 1997;**179**(9):2949-57.

179. Emond E, Holler BJ, Boucher I, Vandenbergh PA, Vedamuthu ER, Kondo JK, Moineau S. Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from Lactococcus lactis. *Appl Environ Microbiol.* 1997;**63**(4):1274-83.

180. Wang C, Villion M, Semper C, Coros C, Moineau S, Zimmerly S. A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA in vitro. *Nucleic Acids Res.* 2011;**39**(17):7620-9.

181. Simon DM, Kelchner SA, Zimmerly S. A broadscale phylogenetic analysis of group II intron RNAs and intron-encoded reverse transcriptases. *Mol Biol Evol.* 2009;**26**(12):2795-808.

182. Doolittle RF, Feng DF, Johnson MS, McClure MA. Origins and evolutionary relationships of retroviruses. *Q Rev Biol.* 1989;**64**(1):1-30.

183. Lambowitz AM, Belfort M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr.* 2015;**3**(1). pii: MDNA3-0050-2014.

184. Robart AR, Chan RT, Peters JK, Rajashankar KR, Toor N. Crystal structure of a eukaryotic group II intron lariat. *Nature.* 2014;**514**(7521):193-7.

185. Costa M, Walbott H, Monachello D, Westhof E, Michel F. Crystal structures of a group II intron lariat primed for reverse splicing. *Science.* 2016;**354**(6316): aaf9258.

186. Zhao C, Pyle AM. Structural Insights into the Mechanism of Group II Intron Splicing. *Trends Biochem Sci.* 2017;**42**(6):470-482.

187. Cech TR. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell.* 1986;**44**(2):207-10.

188. Cavalier-Smith T. Intron phylogeny: a new hypothesis. *Trends Genet.* 1991;**7**(5):145-8.

189. Villa T, Pleiss JA, Guthrie C. Spliceosomal snRNAs: Mg(2+)-dependent chemistry at the catalytic core?. *Cell.* 2002;**109**(2):149-52.

190. Jacquier A. Self-splicing group II and nuclear pre-mRNA introns: how similar are they?. *Trends Biochem Sci.* 1990;**15**(9):351-4.

191. Madhani HD, Guthrie C. Dynamic RNA-RNA interactions in the spliceosome. *Annu Rev Genet.* 1994;**28**:1-26.

192. Keating KS, Toor N, Perlman PS, Pyle AM. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA.* 2010;**16**(1):1-9.

193. Fica SM, Tuttle N, Novak T, Li NS, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA. RNA catalyses nuclear pre-mRNA splicing. *Nature.* 2013;**14**503(7475):229-34.

194. Galej WP, Nguyen TH, Newman AJ, Nagai K. Structural studies of the spliceosome: zooming into the heart of the machine. *Curr Opin Struct Biol.* 2014;**25**:57-66.

195. Strobel SA. Biochemistry: Metal ghosts in the splicing machine. *Nature.* 2013;**503**(7475):201-2.

196. Toor N, Keating KS, Taylor SD, Pyle AM. Crystal structure of a self-spliced group II intron. *Science.* 2008;**320**(5872):77-82.

197. Novikova O, Belfort M. Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends Genet.* 2018;**33**(11):773-783.

198. Qu G, Kaushal PS, Wang J, Shigematsu H, Piazza CL, Agrawal RK, Belfort M, Wang HW. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol.* 2016;**23**(6):549-57.

199. Zhao C, Pyle AM. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol.* 2016;**23**(6):558-65.

200. Agrawal RK, Wang HW, Belfort M. Forks in the tracks: Group II introns, spliceosomes, telomeres and beyond. *RNA Biol.* 2016;**13**(12):1218-1222.

201. Dlakić M, Mushegian A. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA.* 2011;**17**(5):799-808.

202. Galej WP, Oubridge C, Newman AJ, Nagai K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature.* 2013;**493**(7434):638-43.

203. Leclercq S, Giraud I, Cordaux R. Remarkable abundance and evolution of mobile group II introns in Wolbachia bacterial endosymbionts. *Mol Biol Evol.* 2011;**28**(1):685-97.

204. Catania F, Gao X, Scofield DG. Endogenous mechanisms for the origins of spliceosomal introns. *J Hered.* 2009;**100**(5):591-6.

205. Palmer JD, Logsdon JM Jr. The recent origins of introns. *Curr Opin Genet Dev.* 1991;**1**(4):470-7.

206. Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. *Nature.* 2006;**440**(7080):41-5.

207. Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb Perspect Biol.* 2014;**6**(6). pii: a016071.

208. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol.* 2011;**7**(3):e1002021.

209. Mehmood MA, Sehar U, Ahmad N. Use of Bioinformatics Tools in Different Spheres of Life Sciences. *J Data Mining Genomics Proteomics.* 2014;**5**:158.

210. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;**215**(3):403-10.

211. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007;**23**(21):2947-8.

212. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 2000;**302**(1):205-17.

213. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;**32**(5):1792-7.

214. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 2009;**23**(1):205-11.

215. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 2010;**27**(2):221-4.

216. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. *Bioinformatics.* 2004;**20**(3):426-7.

217. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004;**14**(6):1188-90.

218. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 2003;**31**(13):3406-15.

219. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics.* 2009;**25**(10):1335-7.

220. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015;**10**(6):845-58.

221. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* 2010;**5**(4):725-38.

222. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 2003;**31**(13):3784-8.

223. Felsenstein J. PHYLIP phylogeny inference package. *Department of Genetics, University of Washington, Seattle.* 1993.

224. Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences. American Mathematical Society.* 1987;**17**:57–86.

225. Dayhoff MO, Schwartz RM, Orcutt BC. A model for evolutionary change in proteins. *Atlas of Protein Sequence and Structure.* 1978;**5**:345-352.

226. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;**30**(9):1312-3.

227. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;**59**(3):307-21.

228. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 2012;**61**(3):539-42.

229. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evo.* 2018;**35**(6):1547-1549.

230. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res.* 2007;**36**(Database issue):D25-30.

231. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 2005;**33**(Database issue):D29-33.

232. Miyazaki S, Sugawara H, Gojobori T, Tateno Y. DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.* 2003;**31**(1):13-6.

233. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* 2008;**36**(Database issue):D190-5.

234. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003;**31**(1):365-70.

235. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000;**28**(1):235-42.

236. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Zhang D, Bryant SH. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*2013;**41**(Database issue):D348-52.

237. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. *Nucleic Acids Res.* 2010;**38**(Database issue):D211-22.

238. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res.* 2003;**31**(1):439-41.

239. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Schröder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 2013;**41**(Database issue):D605-12.

240. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 2012;**40**(Database issue):D700-5.

241. Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, Kishore R, Lee R, Li Y, Muller HM, Nakamura C, Ozersky P, Paulini M, Raciti D, Schindelman G, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Wong JD, Yook K, Schedl T, Hodgkin J, Berriman M, Kersey P, Spieth J, Stein L, Sternberg PW. WormBase 2014: new views of curated biology. *Nucleic Acids Res.* 2014;**42**(Database issue):D789-93.

242. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ; FlyBase Consortium. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic Acids Res.* 2016;**44**(D1):D786-92.

243. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res.* 2013;**41**(Database issue):D854-60.

244. Karimi K, Fortriede JD, Lotay VS, Burns KA, Wang DZ, Fisher ME, Pells TJ, James-Zorn C, Wang Y, Ponferrada VG, Chu S, Chaturvedi P, Zorn AM, Vize PD. Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res.* 2018;**46**(D1):D861-D868.

245. Abebe M, Candales MA, Duong A, Hood KS, Li T, Neufeld RAE, Shakenov A, Sun R, Wu L, Jarding AM, Semper C, Zimmerly S. A pipeline of programs for collecting and analyzing group II intron retroelement sequences from GenBank. *Mob DNA.* 2013;**4**(1):28.

246. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;**10**:421.

247. Castro C, Smidansky ED, Arnold JJ, Maksimchuk KR, Moustafa I, Uchida A, Götte M, Konigsberg W, Cameron CE. Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol.* 2009;**16**(2):212-8.

248. Coté ML, Roth MJ. Murine Leukemia Virus Reverse Transcriptase: Structural Comparison with HIV-1 Reverse Transcriptase. *Virus Res.* 2008;**134**(1-2):186-202.

249. Steitz TA. DNA polymerases: structural diversity and common mechanisms. *J Biol Chem.* 1999;**274**(25):17395-8.

250. Mönttinen HA, Ravantti JJ, Stuart DI, Poranen MM. Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases. *Mol Biol Evol.* 2014;**31**(10):2741-52.

251. Brutlag D, Schekman R, Kornberg A. A possible role for RNA polymerase in the initiation of M13 DNA synthesis. *Proc Natl Acad Sci U S A.* 1971;**68**(11):2826-9.

252. Wickner W, Brutlag D, Schekman R, Kornberg A. RNA synthesis initiates in vitro conversion of M13 DNA to its replicative form. *Proc Natl Acad Sci U S A.* 1972;**69**(4):965-9.

253. Scherzinger E, Lanka E, Morelli G, Seiffert D, Yuki A. Bacteriophage-T7-induced DNA-priming protein. A novel enzyme involved in DNA replication. *Eur J Biochem.* 1977;**72**(3):543-58.

254. Rowen L, Kornberg A. Primase, the dnaG protein of Escherichia coli. An enzyme which starts DNA chains. *J Biol Chem.* 1978;**253**(3):758-64.

255. Dai L, Toor N, Olson R, Keeping A, Zimmerly S. Database for mobile group II introns. *Nucleic Acids Res.* 2003;**31**(1):424-6.

256. Candales MA, Duong A, Hood KS, Li T, Neufeld RA, Sun R, McNeil BA, Wu L, Jarding AM, Zimmerly S. Database for bacterial group II introns. *Nucleic Acids Res.* 2012;**40**(Database issue):D187-90.

257. Michel F, Costa M, Doucet AJ, Ferat JL. Specialized lineages of bacterial group II introns. *Biochimie.* 2007;**89**(4):542-53.

258. Tourasse NJ, Stabell FB, Kolstø AB. Diversity, mobility, and structural and functional evolution of group II introns carrying an unusual 3' extension. *BMC Res Notes.* 2011;**4**:564.

259. Toor N, Zimmerly S. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. *RNA.* 2002;**8**(11):1373-7.

260. Salman V, Amann R, Shub DA, Schulz-Vogt HN. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc Natl Acad Sci U S A.* 2012;**109**(11):4203-8.

261. Mrázek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A.* 1998;**95**(7):3720-5.

262. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol.* 1996;**13**(5):660-5.

263. Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci U S A.* 2005;**102**(28):9836-41.

264. Rest JS, Mindell DP. Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol.* 2003;**20**(7):1134-42.

265. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glöckner FO, Rosselló-Móra R. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol.* 2008;**31**(4):241-50.

266. Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO, Rosselló-Móra R. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst Appl Microbiol.* 2010;**33**(6):291-9.

267. Cousineau B, Lawrence S, Smith D, Belfort M. Retrotransposition of a bacterial group II intron. *Nature.* 2000;**404**(6781):1018-21.

268. Gao F, Luo H, Zhang CT. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.* 2013;**41**(Database issue):D90-3.

269. Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. Atypical at skew in Firmicute genomes results from selection and not from mutation. *PLoS Genet.* 2011;**7**(9):e1002283.

270. Parikh H, Singh A, Krishnamachari A, Shah K. Computational prediction of origin of replication in bacterial genomes using correlated entropy measure (CEM). *Biosystems.* 2015;**128**:19-25.

271. Pfreundt U, Hess WR. Sequential splicing of a group II twintron in the marine cyanobacterium Trichodesmium. *Sci Rep.* 2015;**5**:16829.

272. Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA.* 2008;**14**(9):1704-13.

273. Qin PZ, Pyle AM. Stopped-flow fluorescence spectroscopy of a group II intron ribozyme reveals that domain 1 is an independent folding unit with a requirement for specific Mg2+ ions in the tertiary structure. *Biochemistry.* 1997;**36**(16):4718-30.

274. Costa M, Déme E, Jacquier A, Michel F. Multiple tertiary interactions involving domain II of group II self-splicing introns. *J Mol Biol.* 1997;**267**(3):520-36.

275. Chanfreau G, Jacquier A. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J.* 1996;**15**(13):3466-76.

276. Fedorova O, Mitros T, Pyle AM. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol.* 2003;**330**(2):197-209.

277. Fedorova O, Pyle AM. Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *EMBO J.* 2005;**24**(22):3906-16.

278. Michel F, Umesono K, Ozeki H. Comparative and functional anatomy of group II catalytic introns–a review. *Gene.* 1989;**82**(1):5-30.

279. Chu VT, Adamidi C, Liu Q, Perlman PS, Pyle AM. Control of branch-site choice by a group II intron. *EMBO J.* 2001;**20**(23):6866-76.

280. Gaur RK, McLaughlin LW, Green MR. Functional group substitutions of the branchpoint adenosine in a nuclear pre-mRNA and a group II intron. *RNA.* 1997;**3**(8):861-9.

281. Toor N, Hausner G, Zimmerly S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA.* 2001;**7**(8):1142-52.

282. Toro N, Molina-Sánchez MD, Fernández-López M. Identification and characterization of bacterial class E group II introns. *Gene.* 2002;**299**(1-2):245-50.

283. Fontaine JM, Goux D, Kloareg B, Loiseaux-de Goër S. The reverse-transcriptase-like proteins encoded by group II introns in the mitochondrial genome of the brown alga Pylaiella littoralis belong to two different lineages which apparently coevolved with the group II ribosyme lineages. *J Mol Evol.*1997;**44**(1):33-42.

284. Kelchner SA. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Am J Bot.* 2002;**89**(10):1651-69.

285. Savill NJ, Hoyle DC, Higgs PG. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics.* 2001;**157**(1):399-411.

286. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* 2011;**27**(8):1164-5.

287. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 2012;**9**(8):772.

288. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics.* 2001;**17**(12):1246-7.

289. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;**24**(8):1586-91.

290. Van Roey P, Meehan L, Kowalski JC, Belfort M, Derbyshire V. Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat Struct Biol.* 2002;**9**(11):806-11.

291. Nimkulrat S, Lee H, Doak TG, Ye Y. Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with Treponema denticola. *Front Microbiol.* 2016;**7**:852.

292. Naorem SS, Han J, Wang S, Lee WR, Heng X, Miller JF, Guo H. DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. *Proc Natl Acad Sci U S A.* 2017;**114**(47):E10187-E10195.

293. Guo H, Arambula D, Ghosh P, Miller JF. Diversity-generating Retroelements in Phage and Bacterial Genomes. *Microbiol Spectr.* 2014;**2**:(6).

294. Guo H, Tse LV, Nieh AW, Czornyj E, Williams S, Oukil S, Liu VB, Miller JF. Target site recognition by a diversity-generating retroelement. *PLoS Genet.* 2011;**7**(12):e1002414.

295. Park J, Zhang Y, Buboltz AM, Zhang X, Schuster SC, Ahuja U, Liu M, Miller JF, Sebaihia M, Bentley SD, Parkhill J, Harvill ET. Comparative genomics of the classical Bordetella subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics.* 2012;**13**:545.

296. Paul BG, Bagby SC, Czornyj E, Arambula D, Handa S, Sczyrba A, Ghosh P, Miller JF, Valentine DL. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nat Commun.* 2015;**6**:6585.

297. Schillinger T, Zingler N. The low incidence of diversity-generating retroelements in sequenced genomes. *Mob Genet Elements.* 2012;**2**(6):287-291.

298. Ye Y. Identification of diversity-generating retroelements in human microbiomes. *Int J Mol Sci.* 2014;**15**(8):14234-46.

299. Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, Czornyj E, Thomas BC, Ghosh P, Miller JF, Banfield JF, Valentine DL. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat Microbiol.* 2017;**2**:17045.

300. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature.* 2015;**523**(7559):208-11.

301. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome.* 2017;**5**(1):112.

302. Dimmic MW, Rest JS, Mindell DP, Goldstein RA. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol.* 2002;**55**(1):65-73.

303. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 2016;**44**(W1):W16-21.

304. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. Insights into the phylogeny and coding potential of microbial dark matter. *Nature.* 2013;**499**(7459):431-7.

305. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. A new view of the tree of life. *Nat Microbiol.* 2016;**1**:16048.

306. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM, Taylor RC, Williams KH, Banfield JF. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol.* 2015;**25**(6):690-701.

307. Handa S, Paul BG, Miller JF, Valentine DL, Ghosh P. Conservation of the C-type lectin fold for accommodating massive sequence variation in archaeal diversity-generating retroelements. *BMC Struct Biol.* 2016;**16**(1):13.

308. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016;**44**(W1):W3-W10.

309. Siguier P, Varani A, Perochon J, Chandler M. Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. *Methods Mol Biol.* 2012;**859**:91-103.

310. Neumann P, Pozárková D, Macas J. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. *Plant Mol Biol.* 2003;**53**(3):399-410.

311. Pyatkov KI, Shostak NG, Zelentsova ES, Lyozin GT, Melekhin MI, Finnegan DJ, Kidwell MG, Evgen'ev MB. Penelope retroelements from Drosophila virilis are active after transformation of Drosophila melanogaster. *Proc Natl Acad Sci U S A.* 2002;**99**(25):16150-5.

312. Gillis AJ, Schuller AP, Skordalakes E. Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature.* 2008;**455**(7213):633-7.

313. Mitchell M, Gillis A, Futahashi M, Fujiwara H, Skordalakes E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol.* 2010;**17**(4):513-8.

314. Jaskelioff M, Muller FL, Paik JH, Thomas E, Jiang S, Adams AC, Sahin E, Kost-Alimova M, Protopopov A, Cadiñanos J, Horner JW, Maratos-Flier E, Depinho RA. Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. *Nature.* 2011;**469**(7328):102-6.

315. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct.* 2006;**1**:7.

316. Schmelzer C, Schweyen RJ. Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation. *Cell.* 1986;**46**(4):557-65.

317. Pyle AM. The tertiary structure of group II introns: implications for biological function and evolution. *Crit Rev Biochem Mol Biol.* 2010;**45**(3):215-32.

318. Robart AR, Seo W, Zimmerly S. Insertion of group II intron retroelements after intrinsic transcriptional terminators. *Proc Natl Acad Sci U S A.* 2007;**104**(16):6620-5.

319. Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM. Group II intron-based gene targeting reactions in eukaryotes. *PLoS One.* 2008;**3**(9):e3121.

320. White TB, Lambowitz AM. The retrohoming of linear group II intron RNAs in Drosophila melanogaster occurs by both DNA ligase 4-dependent and -independent mechanisms. *PLoS Genet.* 2012;**8**(2):e1002534.

321. Muñoz-Adelantado E, San Filippo J, Martínez-Abarca F, García-Rodríguez FM, Lambowitz AM, Toro N. Mobility of the Sinorhizobium meliloti group II intron RmInt1 occurs by reverse splicing into DNA, but requires an unknown reverse transcriptase priming mechanism. *J Mol Biol.* 2003;**327**(5):931-43.

322. Sellem CH, Lecellier G, Belcour L. Transposition of a group II intron. *Nature.* 1993;**366**(6451):176-8.

323. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol.* 2003;**4**(11):865-77.

324. Stamos JL, Lentzsch AM, Lambowitz AM. Structure of a Thermostable Group II Intron Reverse Transcriptase with Template-Primer and Its Functional and Evolutionary Implications. *Mol Cell.* 2017;**68**(5):926-939.e4.

325. Hua-Van A, Le Rouzic A, Maisonhaute C, Capy P. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet Genome Res.* 2005;**110**(1-4):426-40.

326. Lampson BC, Sun J, Hsu MY, Vallejo-Ramirez J, Inouye S, Inouye M. Reverse transcriptase in a clinical strain of Escherichia coli: production of branched RNA-linked msDNA. *Science.* 1989;**243**(4894 Pt 1):1033-8.

327. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 2011;**21**(10):1616-25.

Appendices

Appendices A, B, D, E are provided in a zip file and is available on the University of Calgary thesis repository – The Vault (https://prism.ucalgary.ca/).

Appendix A. Supplemental data for Chapter II.

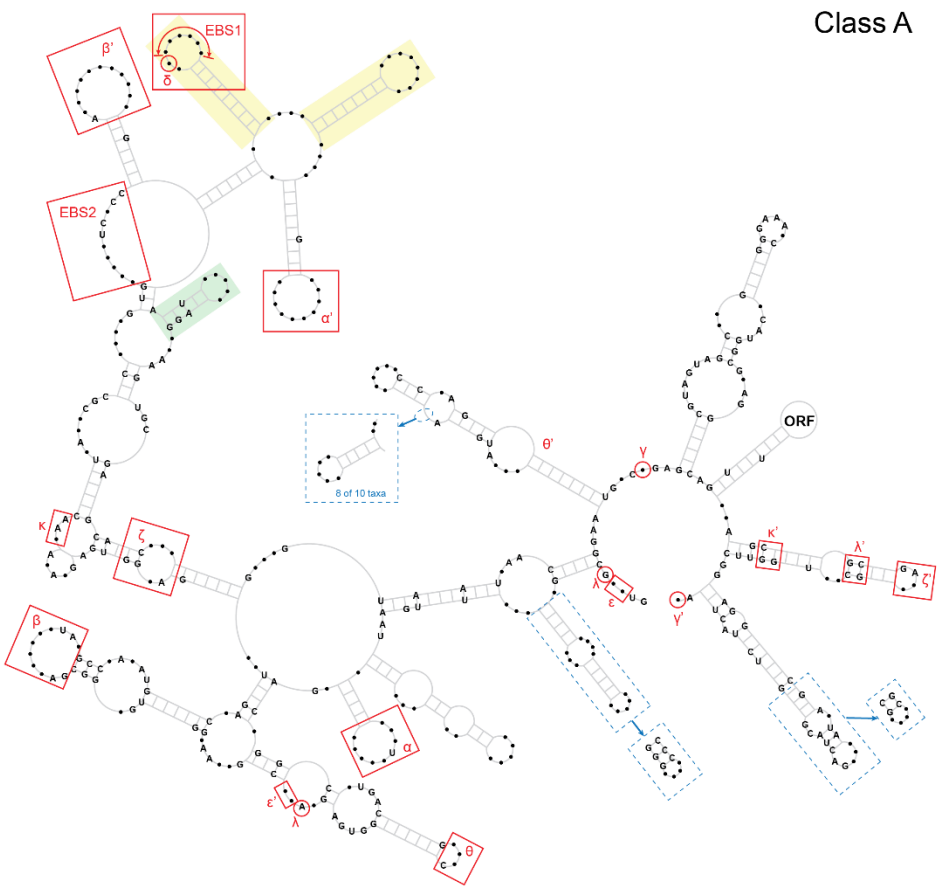
ID	Filename	Description
A1	initial_tblastn_query.fst	Initial BLAST query sequences used at the beginning step of TBLASTN in FASTA format.
A2	custom_rt_db.7z	Custom database for RT classification. Both original sequences and the BLAST database generated by NCBI's BLAST+ suite are provided. The type of each sequence is included at the end of the name.

Appendix B. Supplemental data for Chapter III.

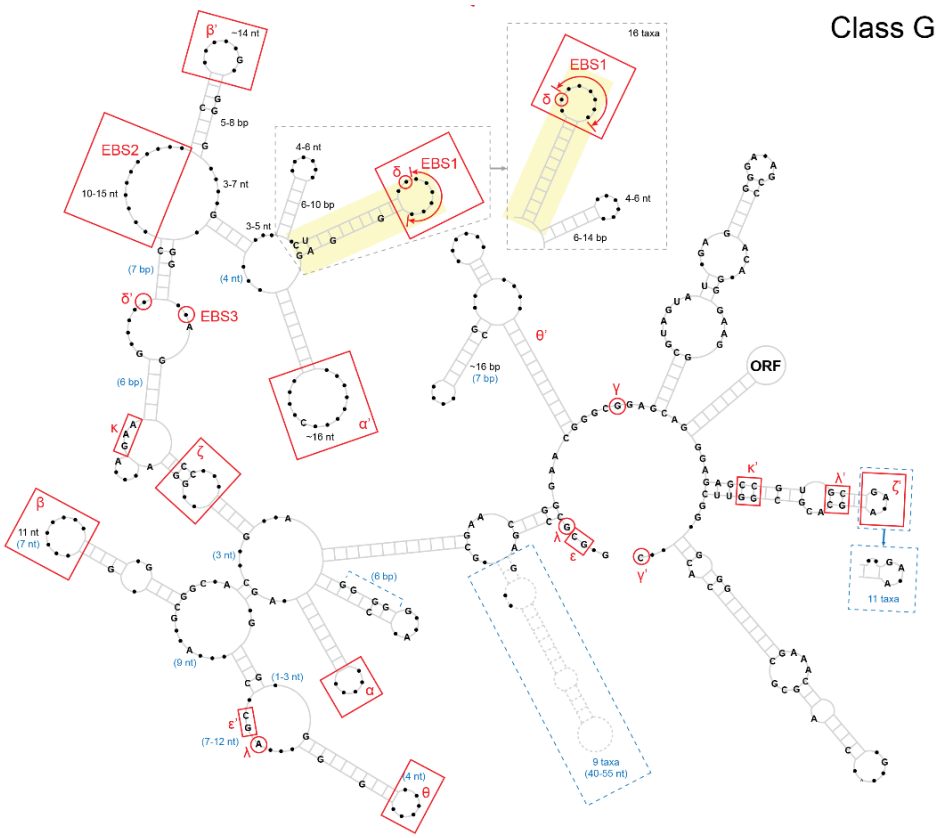
ID	Filename	Description
B1	domain_map .fst	Domain map of all Classes A to ML. The alignment includes one representative sequences for each class. A sequence named "mask" is placed on top, which indicates selected positions within all RT motifs 0-7, as well as domains X and EN (if present). Digits "0-7" correspond to RT motifs 0-7, "a" corresponds to RT motif 2a, "x" corresponds to domain X, and "E" corresponds to domain EN (only applicable to Classes B, CL1, CL2 and ML).
B2	tree_by_phyla.png; tree_by_phyla_key.png	The IEP-based phylogenetic tree of 1275 introns, colour coded by their phyla. Tree was constructed using RAxML with 1000 bootstrap replicates. Supported nodes (bootstrap >= 75%) are indicated by a black dot. The class of each intron is noted at the end of the intron name. Due to the large size of this figure, key colours are provided separately as "tree_by_phyla_key.png".
B3	list_twintan.xlsx	List of unique twintron and tandem intron units.
B4	genome_map.7z	Genomic mapping of group II intron and transposon-related elements in three sample genomes: CP013008 (*Arthrospira platensis YZ),* FO818640 (*Arthrospira sp. str. PCC 8005) and* CP000393 (*Trichodesmium erythraeum IMS101).* The whole genome was searched for group II RNA and IEP segments using HMMER, and transposon-related elements were searched using the ISfinder web server. Figures are drawn to scale in SVG format. Hits to different elements are colour coded as below: Red: a match to domains I, II and III of the group II ribozyme; Blue: a match of domains V and VI of the group II ribozyme; Purple: a match to both 5'- and 3'-portions of the group II ribozyme, usually happens when the two portions of RNA overlap; Green: a match to the group II IEP; Light green: a match to any transposon-related element. Yellow shading: the corresponding intron is already recorded in our group II database.
B5	db_stru.xlsx	Description of fields for all tables in the order as labelled in Figure 24.

Appendix C. Consensus structures of Classes A, G and H.

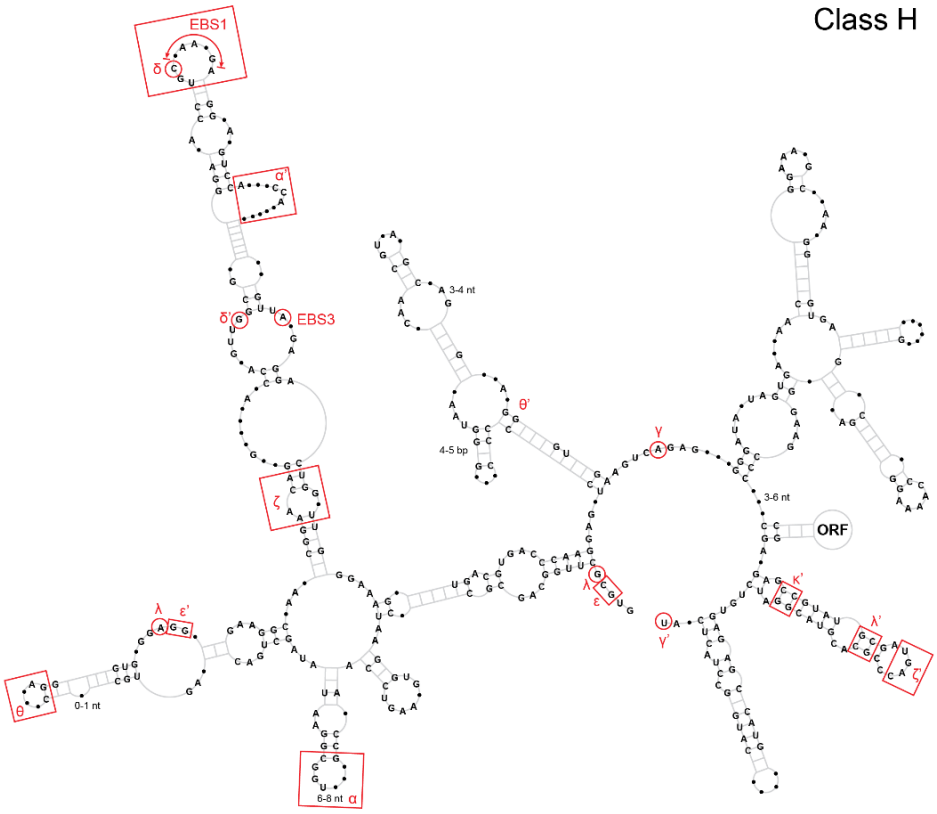
The consensus sequence was created from the structural RNA alignment, using a sequence identity threshold of 100%, 90% and 100% for Classes A, G and H respectively. Major tertiary interactions are shown in red circles or boxes. Areas in blue dashed boxes show alternative structures conserved in specific subsets. Classes A and G were noticed to have two potential stem-loops for EBS1, which are shaded in yellow. For Class A, the ID(iii)2-like structure is shaded in green, which is not found in other classes except for ML.



Class G



Class H



Appendix D. Supplemental data for Chapter IV.

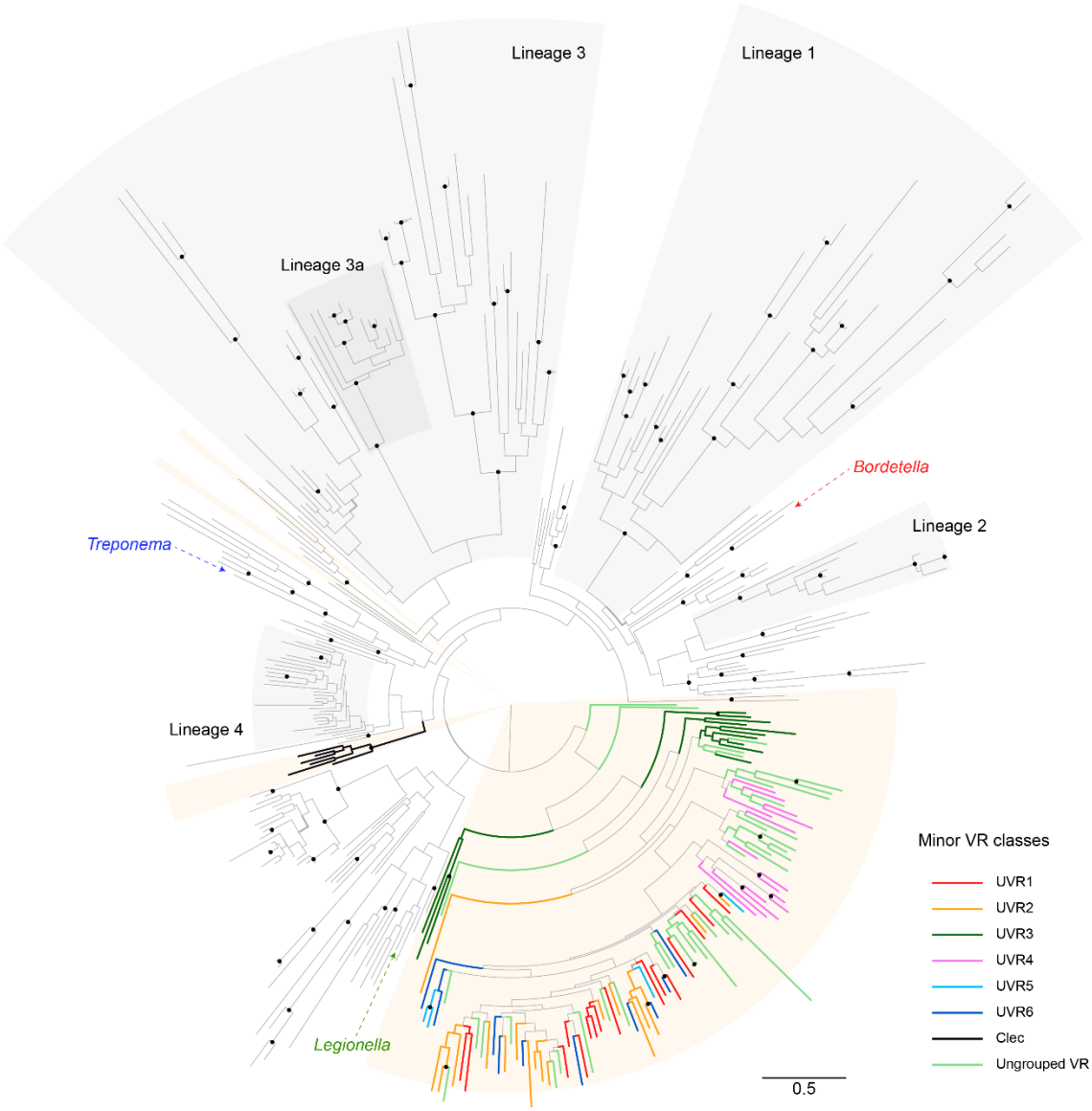
ID	Filename	Description
D1	taxalist.xlsx	List of taxa used in this project. Worksheet "intron_taxa" lists all group II introns. The "GC-IEP" and "GC-RNA" columns give the GC content of ranges of IEP only and the entire RNA. Columns "taxon 1-4" correspond to the four subsets for taxon sampling, column "GC bias" corresponds to the GC-content based sampling, and column "external" corresponds to the representatives used while being aligned to external RTs. Worksheet "external" lists all non-group II sequences. The source of sequence is listed in column "Database/Source", which is either a sequence database or a research article (provided as doi). The column "Accession" indicates either the accession number if the sequence is in a database, or the sequence name if it was from a research article. The column "Range" is only applicable to sequences from NCBI's nucleotide database, the start and the end is separated by a double-dot "..", and parentheses indicate the sequence is on the complementary strand.
D2	morphology.xlsx	Morphological characters for each class and unclassified introns. Each class was treated as a single taxon, and all six unclassified introns were included.
D3	sh_test.xlsx	Results of SH topology tests. The SH test was performed within each class by programs PAML and CONSEL. The input tree list included all class-specific trees as well as subtrees extracted from global trees, which was compared individually to class-specific alignments after applying every sequence mask. The significance level was set to $\alpha = 0.05$. A plus sign "+" indicates values higher than the significance level (pass the SH test). "N/A" indicates the corresponding mask was not available.

Appendix E. Supplemental data for Chapter V.

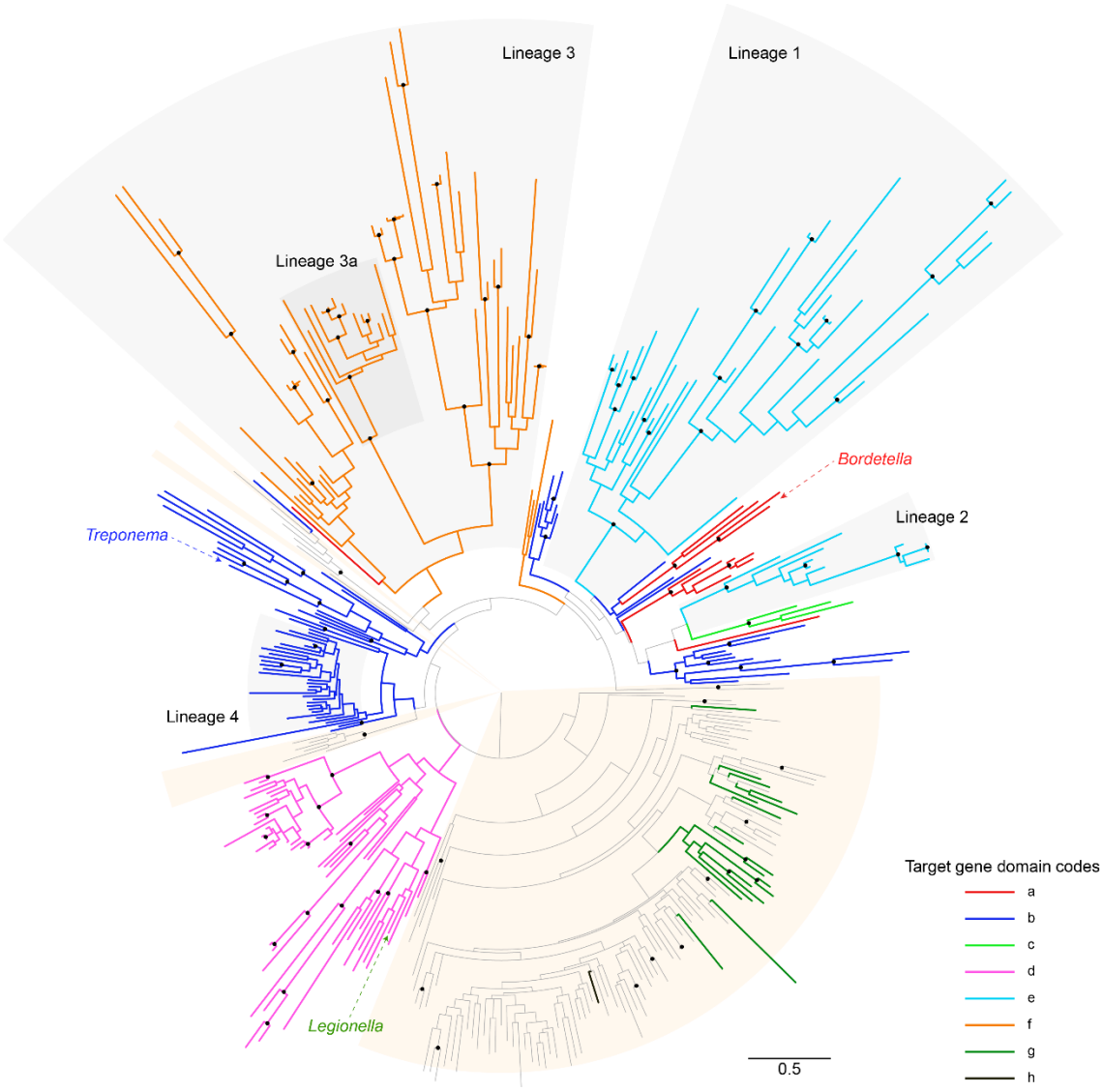
ID	Filename	Description
E1	seeds_for_update.txt	Query sequences used in the second update in 2016. Both the new IDs and old identifiers are given.
E2	dgr_master.xlsx	Master table of the 372 unique DGRs. "ST1" contains the compiled information of DGRs, "ST1b" contains descriptions of each column in "ST1". "ST2" contains informations of ORFs between known DGR components. "ST3" contains description of domains grouped by the TG category. This is the same file as provided to the publication.
E3	tr_vr.7z	TR-VR alignments of DGR. Alignments are organised by the VR class in HTML format. Both the VR and TR sequences as well as their corresponding proteins are provided. Canonical A-to-N mismatches are highlighted in yellow, and other mismatches and indels are highlighted in green. This is the same file as provided to the publication.
E4	stemloop.tsv.txt	Predicted stem-loops. The boundary of the stem-loop, the folding direction given in Vienna format and the distance to the VR are given. In the "eval" column, "2" and "1" correspond to "high" and "low" confidence respectively. Sequences without a potential stem-loop are omitted.

Appendix F. Additional figures showing the correlation between selected factors and the RT-based phylogenetic tree.

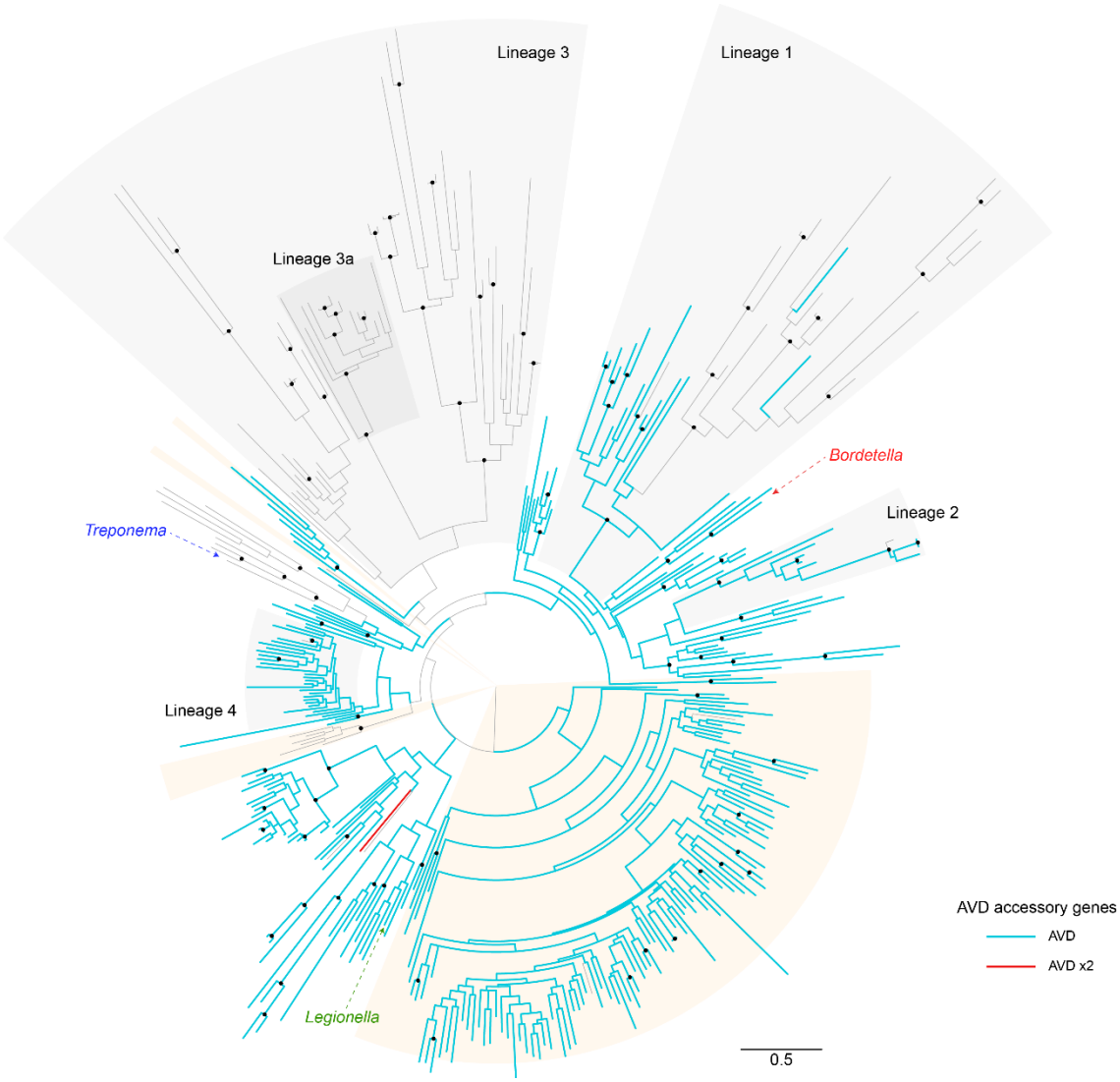
F1. Minor VR classes



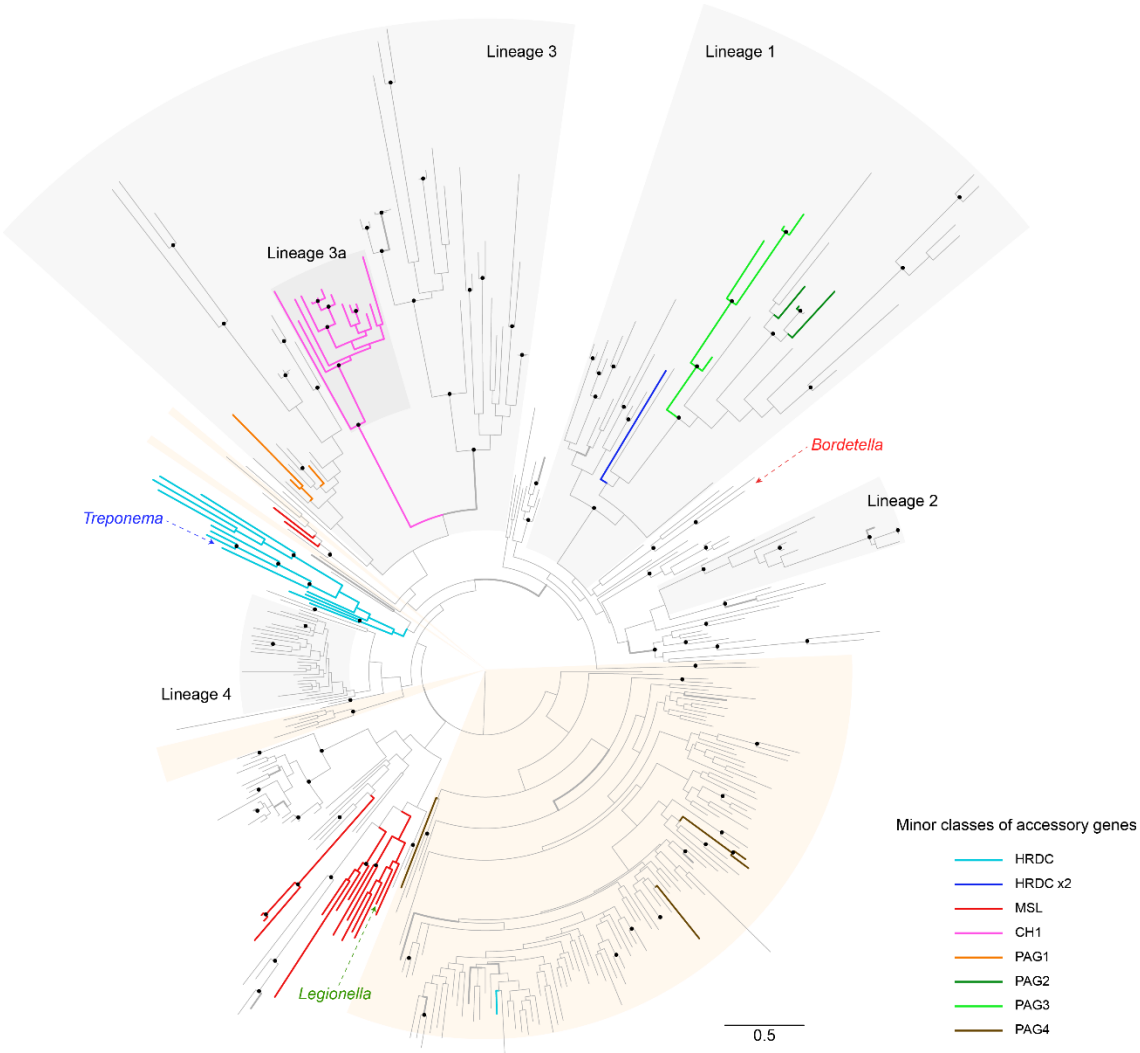
F2. TG domain composition



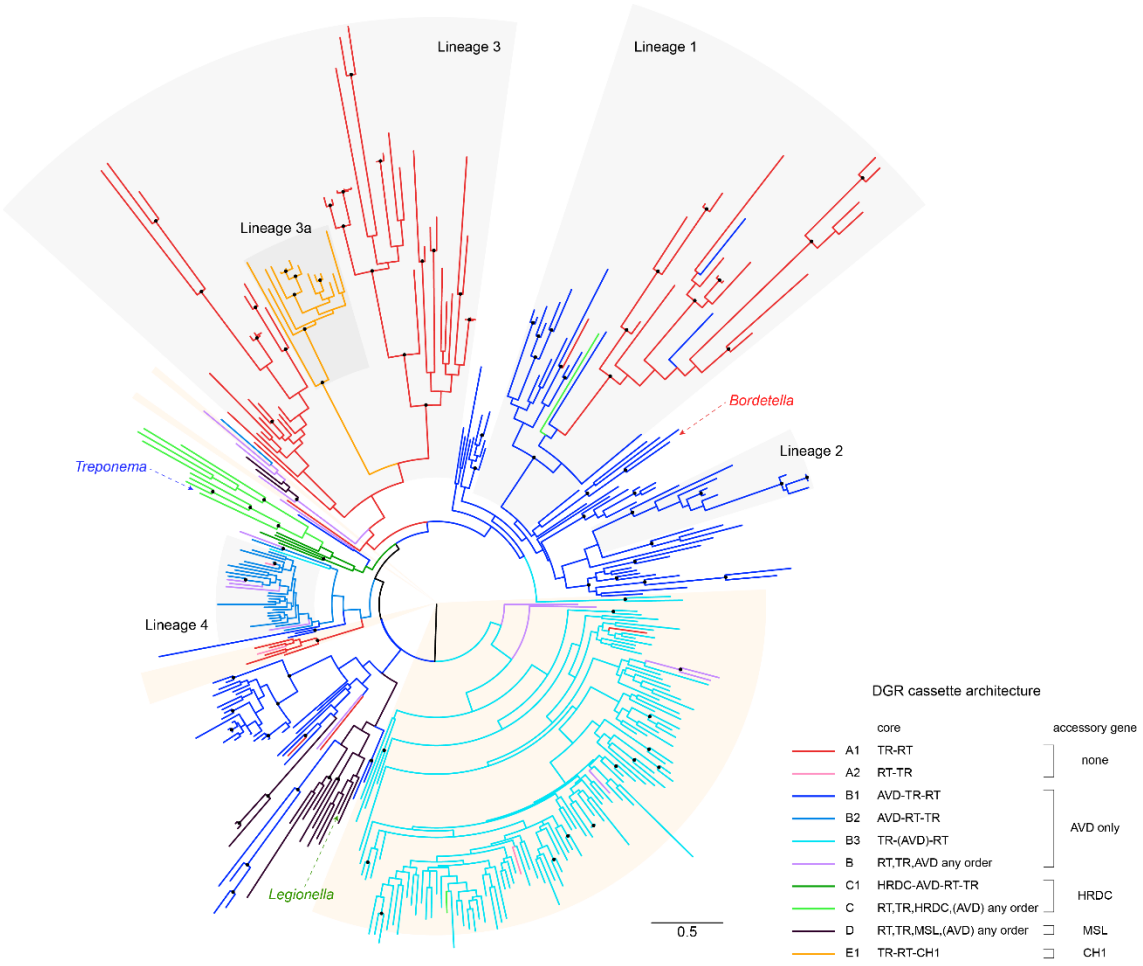
F3. The accessory gene AVD



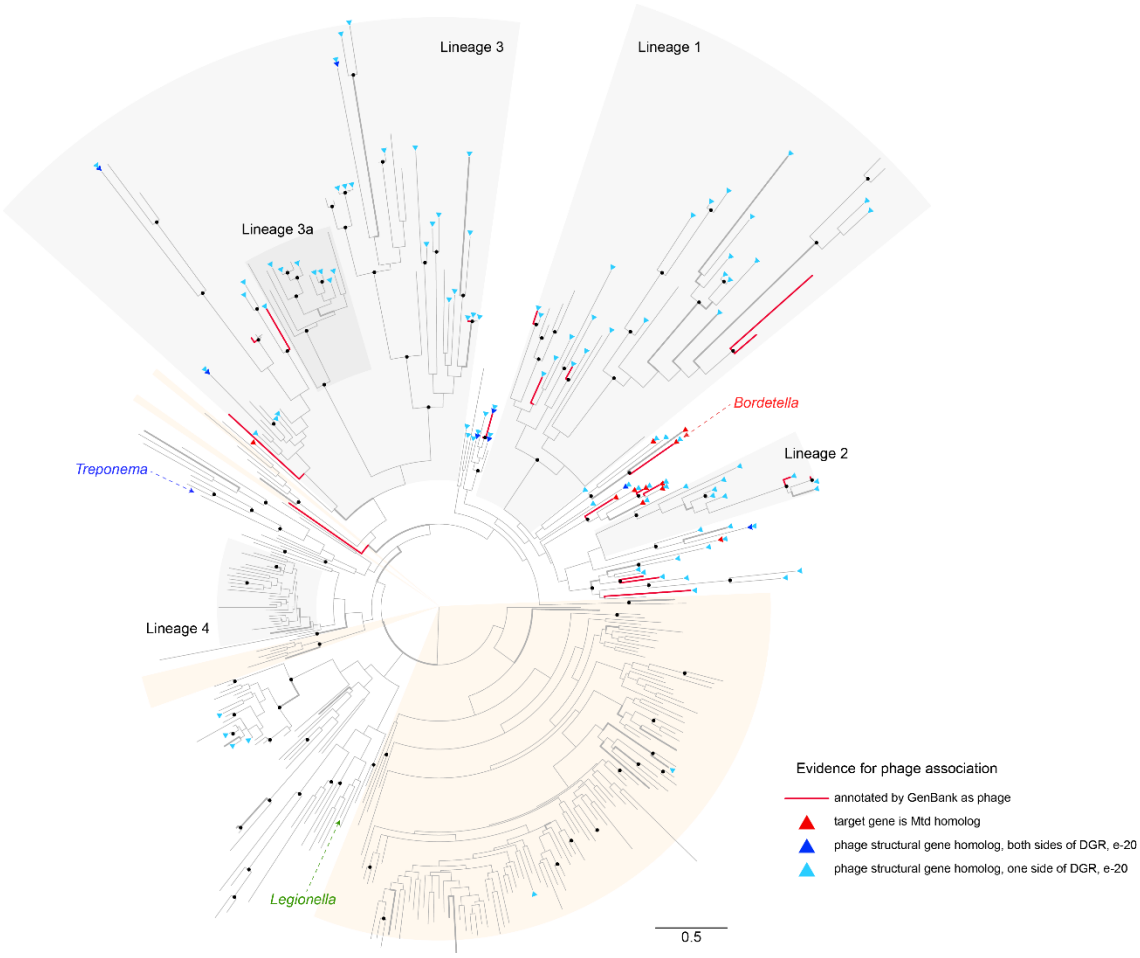
F4. Other minor accessory genes



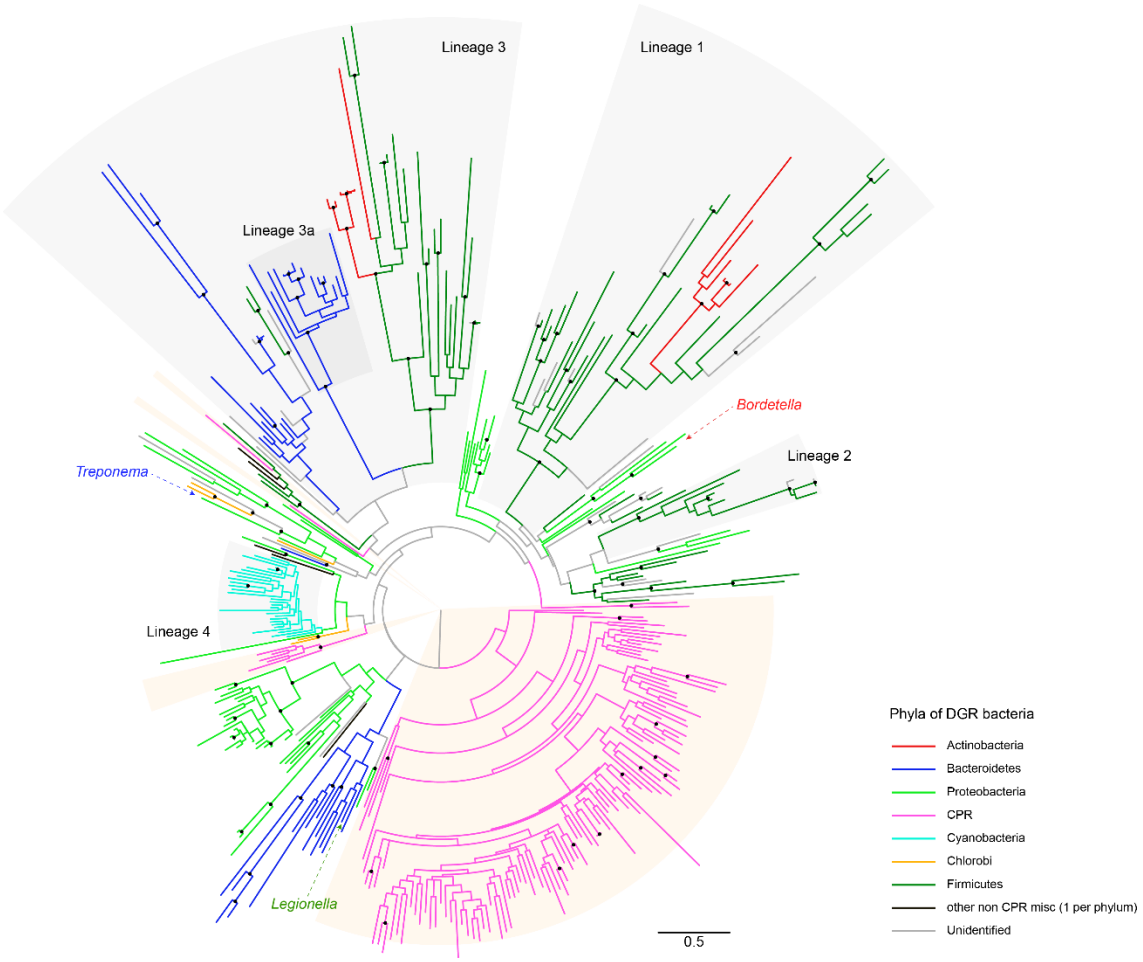
F5. Architectures



F6. Phage association



F7. Phyla of host organisms



F8. Position of VR in TG