# Development of a SELEX method to uncover auto-aminoacylating ribozymes and analysis of aminoacyl RNA from Escherichia coli transcriptomes

Ji Wang

NNT: 2016SACLS269

# THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD

LABORATOIRE
SEQUENCE, STRUCTURE ET FONCTION DES ARN (SSFA), I2BC

ECOLE DOCTORALE N° 577

Structure et dynamique des systèmes vivants (SDSV)

Spécialité de doctorat : Sciences de la Vie et de la Santé
Biologie moléculaire

Par

**M. Ji Wang**

Développement d'une méthode SELEX pour l'identification de ribozymes pour l'aminoacylation et analyse d'ARN aminoacylés dans le transcriptome d'*Escherichia coli*

**Thèse présentée et soutenue à ORSAY, le date : 16/09/2016**

**Composition du Jury :**

M., Christophe SOLA, Professeur, Université Paris-Sud, Président
Mme, Sabine MüLLER, Professeure, Ernst-Moritz-Arndt University Greifswald, Rapporteur
M., Benoît MASQUIDA, Directeur de recherche, CNRS Strasbourg, Rapporteur
Mme, Tamara BASTA-LE BERRE, Maître de Conférences, Université Paris-Sud, Examinatrice
M., Fabrice JOSSINET, Maître de Conférences, Université de Strasbourg, Examinateur
M., Jean LEHMANN, Maître de Conférences, Université Paris-Sud, Directeur de these

# SYNTHÈSE

## Développement d'une méthode SELEX pour l'identification de ribozymes pour l'aminoacylation et analyse d'ARN aminoacylés dans le transcriptome d'*Escherichia coli*

Les ribozymes sont des ARN naturels ou artificiels possédant une activité catalytique. Les ribozymes artificiels ont été identifiés *in vitro* par la méthode SELEX, et plusieurs d'entre eux ont été caractérisés par des études cinétiques. Ces molécules sont impliquées dans des réactions de clivage, de ligation, de modification d'extrémités d'ARN, de polymérisation, de phosphorylation et d'activation de groupements acyl. Parce qu'elle est nécessaire à la traduction, l'aminoacylation des ARN joue un rôle évolutif important dans la transition du monde de l'ARN vers le monde moderne de l'ADN et des protéines, et elle est centrale à l'établissement du code génétique. Plusieurs ribozymes catalysant le transfert d'acides aminés à partir de cofacteurs activants ont pu être isolés et caractérisés depuis une vingtaine d'années, ce qui a documenté la possibilité d'aminoacylation d'ARNt en l'absence des aminoacyl ARNt synthétases.

En développant un nouveau protocole SELEX basé sur l'oxydation au periodate, le but de notre travail est de découvrir de nouveau ribozymes d'une taille de l'ordre d'une vingtaine de nucléotides pouvant combiner la catalyse de l'activation des acides aminés et la transestérification. Bien que des molécules catalysant l'une ou l'autre des deux réactions ont été identifiées, aucun ribozyme n'existe à ce jour qui puisse utiliser des acides aminés libres et un cofacteur activant pour réaliser l'aminoacylation en 3' dans un même milieu réactionnel.

La sélection de molécules actives dans une approche SELEX exige la présence de régions constantes sur les deux extrémités des séquences pools aléatoires initiaux. Ces régions sont nécessaires pour l'amplification par PCR, mais elles imposent des contraintes importantes pour l'identification de ribozymes car elles peuvent

complètement inhiber leur activité par interférence structurelle. Nous présentons un protocole optimisé qui minimise la taille de ces régions constantes. D'autre part, notre nouveau design est très spécifique pour la sélection d'ARN aminoacylés sur l'extrémité 3'. Ce protocole a été utilisé pour réaliser 6 à 7 cycles de sélection avec différents pools, et un enrichissement en séquences spécifiques a pu être mis en évidence. Bien que certains tests avec les pools sélectionnés aient révélé une activité possible, des essais avec des séquences spécifiques de ces pools n'ont pour l'instant pas pu confirmer l'activité catalytique recherchée.

Un protocole basé sur le même principe de sélection a été utilisé dans une étude parallèle pour identifier les ARN aminoacylés présents dans l'ARN total d'*Escherichia coli*. Dans ce deuxième travail, notre but est d'identifier tous les d'ARN aminoacylés par séquençage massif, avec à la clé la découverte possible de molécules autres que les ARNt et ARNtm. En utilisant les ARNt comme modèle, nous nous sommes aperçus qu'un protocole RNAseq standard n'était pas adapté à cause des bases modifiées présentes sur ces molécules. Nous avons développé et mis au point un nouveau protocole pour l'identification de n'importe quelle séquence aminoacylée en 3'. La nouvelle approche présentée devrait permettre l'étude exhaustive de l'aminoacylation de toutes les séquences présentes dans l'ARN total.

# Summary


**Development of a SELEX method to uncover auto-aminoacylating ribozymes and analysis of aminoacyl RNA from *Escherichia coli* transcriptomes**

Ribozymes are natural or in vitro selected RNA molecules possessing a catalytic activity. Artificial ribozymes have been extensively investigated by *in vitro* SELEX experiments, and characterized by kinetic assays. Ribozymes are involved in RNA cleavage, ligation, capping, polymerization, phosphorylation and acyl activation. Because it is required for translation, RNA aminoacylation plays an important role in the evolution from the late RNA world to the modern DNA and protein world, and is central to the genetic code. Several ribozymes catalyzing amino acid transfer from various activating groups have already been selected and characterized in the past two decades, documenting the possibility of tRNA aminoacylation in the absence of aminoacyl tRNA synthetase.

With a newly designed SELEX protocol based on periodate oxidation, the aim of our investigation is to uncover small ribozymes of the order of 20 nucleotides that could catalyze both amino acid activation and transesterification. Although molecules catalyzing either reaction have been identified, no existing ribozyme could use free amino acids and activating cofactor(s) as substrates for 3' esterification in a single reactional context.

The selection of active molecules in a SELEX procedure requires the presence of constant tracks on both ends of the sequences constituting the initial random pools. These tracks are required for PCR amplification, but they impose significant burden to the identification of ribozymes because they can prevent any activity through structural inhibition. We present an optimized protocol that significantly minimizes the size of these constant tracks. At the same time, our newly design protocol is very specific for the selection of 3'-end aminoacylated RNA. Working with this protocol, we performed

6 to 7 cycles of selection with different pools, and observed an enrichment with specific sequences. Although some experiments performed with entire pools did reveal a possible activity, no activity could be so far confirmed with specific sequences.

A similar protocol was also applied in a parallel study to identify aminoacylated RNA from total RNA in *Escherichia coli*. In this other approach, our goal is to possibly identify new classes of aminoacylated RNA while using the deep sequencing technology. Using tRNA to validate our protocol, we realized that a standard RNAseq procedure could not work due to the presence of modified bases. We established a new method for bank preparation to identify any sequence aminoacylated at the 3' end. Ultimately, this new approach will allow us to study the level of aminoacylation of any sequence present in total RNA.

# ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisor **Jean Lehmann** for his dedicated guidance, sympathy, generosity and supports during all four years. It is him who always encouraged and helped me to overcome the difficult moments in research. I feel extremely lucky to have chance to work with and learn from him. It is my great honor to have opportunity of working in **Daniel Gautheret**'s lab with all my colleagues.

It was also an honor to have collaboration with **Philippe Bouloc**, **Annick Jacq** and my tutor **François Michel**, from whom I received a lot of encouragements and advice.

I am highly thankful to my colleague **Florence Lorieux** who helped me to accomplish the sequencing experiments, and also **Coline Billerey** and **Marc Gabriel** who have contributed to my thesis with bioinformatics analyses.

I would like to thank all my wonderful and kind colleagues **Claire Toffano-Nioche**, **Fabrice Leclerc**, **Nicolas Chevrollier**, **Chantal Bohn**, **Elena Disconzi**, **Rémy Bonnin**, **Claire Morvan** and **Aurélie Jaffrenou**. They have helped me a lot in many occasions since I arrived in France.

I truly appreciate all the helps from my best friends, **Audrey Vingadassalon**, **Lê Lâm Thảo Nguyên**, and **Nguyễn Ngọc Ẩn**. With their whole-hearted supports and kindness, we have been through very memorable experiences together.

I would also like to thank my Chinese colleagues **DENG Yiqin**, **LIU Wenfeng**, **LUO Xing** and **LI Jia**. We have been sharing lots of wonderful moments in daily life, research experiences and making progress together. I especially would like to express my gratitude to **Dr. Shixin Ye-Lehmann** who is always following my work, considering about me and trying the best to help me.

I am grateful for being supported by ***Chinese Scholarship Council*** (**CSC**) [File No. 201206140111] during four years of PhD study and all the hard work from the "Education service office, Embassy of P.R. China in France".

In the end, I would like to thank my parents, **YAO Ruling** and **WANG Feng**. During these four years far away from China, I feel sorry that I cannot be with them so often. Without their comprehensive and unconditional supports, it is impossible for me to finish my study.

I wish all the best to the colleagues from ***Institut de Biologie Intégrative de la Cellule*** (**I2BC**) and appreciate all the helps from them.

*“Trouver n’est rien, c’est le plan qui est difficile.”*

—— *Fiodor Dostoïevski*

# ABBREVIATIONS

| | |
|---|---|
| aa/ aa-AMP | amino acid(s)/ adenylated amino acid(s) |
| aaRS | aminoacyl tRNA synthetase |
| Ala | alanine |
| ATP/AMP | adenosine triphosphate/adenosine monophosphate |
| bp | base pair |
| ds | double strand |
| DTT | dithiothreitol |
| eq. | equation |
| Gly | glycine |
| GNA | glycerol-derived nucleic-acid analogue |
| GTP | guanosine triphosphate |
| HDV | hepatitis delta virus |
| HH | hammerhead |
| His | histidine |
| M-MLV | Moloney murine leukemia virus |
| NAD | nicotinamide adenine dinucleotide |
| nt | nucleotide |
| NTP | nucleoside triphosphate |
| OAc | $CH_3COO^-$ |
| PCR/ RT-PCR | polymerase chain reaction/ reverse transcription - PCR |
| PDB | phosphodiester backbones |
| PEG | polyethylene glycol |
| Phe-AMP | adenylated phenylalanine |
| PNA | peptide nucleic acid |
| PPi | pyrophosphate |
| PVA/PVP | polyvinyl alcohol/polyvinyl pyrrolidine |
| RACE | rapid amplification of cDNA ends |
| Rnl1 | RNA ligase 1 |
| RNP | ribonucleoprotein |
| rRNA | ribosome RNA |
| RT | reverse transcriptase |
| SELEX | systematic evolution of ligands by exponential amplification |
| ss | single-stranded |
| SSIII | SuperScript® III |
| TdT | terminal deoxynucleotidyl transferase |
| tRNA/tmRNA | transfer RNA/transfer-messenger RNA |
| tsRNA/tRFs | tRNA derived small RNA fragments |
| TNA | threose nucleic acid |
| Tyr | tyrosine |
| Val | valine |

# CONTENTS

# Development of a SELEX method to uncover auto-aminoacylating ribozymes and analysis of aminoacylated RNA from *Escherichia coli* transcriptomes

## General Introduction

### 1. Life at the origin

The evolution of the early forms of life is a fundamental problem investigated by scientists that still has many unresolved aspects. All existing life known so far in the "modern biology world" is based on three types of biomolecules: DNA, RNA and protein. It is certainly not straightforward to infer the composition or the metabolism of the primary form(s) of life. There is some kind of agreement that the original life rooted in the evolution of complex chemical mixtures (**Figure 1**) (Grossenbacher and Knight 1965; Oro 2002). Biochemists have already proposed several hypotheses to explain how a "Spontaneous generation" could happen (Balme 1962) and how the organic compounds constituting the original chemical metabolism in the "Primordial soup" came about (Shapiro 1987).

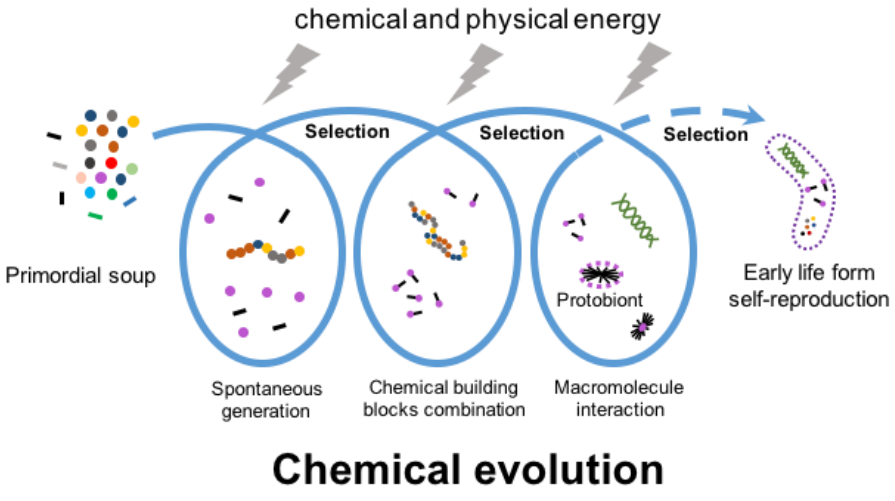



**Figure 1**. Chemical evolution has been undergoing several stages. In the first stage, simple organic substances, such as amino acids, nucleic acids, fatty acids, and sugar, start to form the primordial soup (Bailey 1938). Then the simple organic molecules begin to gathering together to yield larger macromolecules. As the protobiont evolved, it gained the ability of reproduction and delivering the genetic information (Morowitz et al. 1988). Finally, some of the polymers, such as RNA, DNA, peptide, are able to catalyze enzymatic reactions, construct the protocell and build the early life forms (Chen et al. 2005).

During chemical evolution, it has been hypothesized that different molecules existing in the ancient ocean have all been through a natural selection based on their chemical properties (Schwartz 1971). RNA and DNA are not the only polymers to be evolutionary candidates. Hypothetical biology building blocks, such as threose nucleic acid (TNA), peptide nucleic acid (PNA), glycerol-derived nucleic-acid analogue (GNA), (**Figure 2**) share a similar chemical structural profile and are able to form stable Watson–Crick base pairs just like RNA and DNA (Joyce 2002). Furthermore, self-replication could also happen with certain peptides and small organic compounds without "Darwin evolution" (Tjivikua et al. 1990; Lee et al. 1996; Murtas 2013).

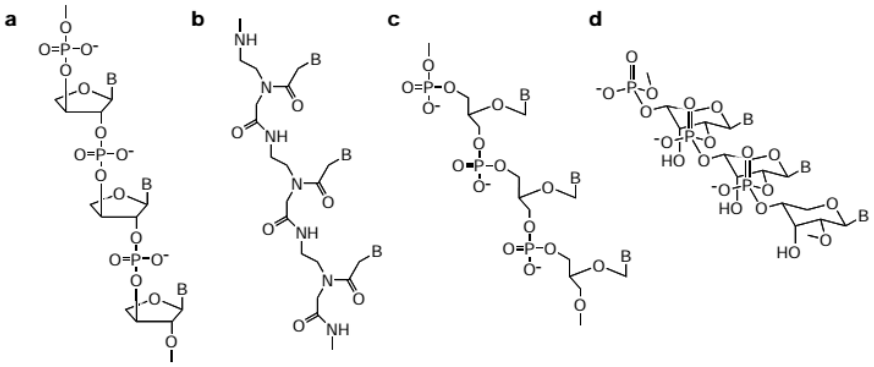



**Figure 2**. Candidate precursors (nucleic acid-like molecules) to nucleic acid. **a**, Threose nucleic acid (TNA); **b**, peptide nucleic acid (PNA); **c**, glycerol-derived nucleic-acid analogue (GNA); **d**, pyranosyl-RNA. "B", nucleotide base. (Joyce 2002)

Thus, information storage and self-replication are not the specialty of nucleic acid only. Nevertheless, RNA and DNA have some advantages over other elementary building blocks that prevailed. Nucleic acids are polymers made up of 4 different types of elementary units which can provide a high diversity of sequences, allow them to be replicated faithfully (compared to peptides), and provide the basis for Darwinian evolution (Joyce 1989). Hypothetically, the original instability of individual RNA or DNA molecules and a low replication efficiency would make very long polymers impossible, and 50-100 bases have been estimated as an upper limit (Eigen and Schuster 1978). However, at some point, an improved stability could promote the assemblage of larger molecules (Yakhnin 2013). Being enriched in electron donors, such as purine and pyrimidine groups, phosphate group and oxygen group, long RNA molecules may

better position the ions binding location and form the initial enzymatic catalysis system (Ralser 2014).

Unlike DNA, single stranded RNA molecules can fold into extraordinary complicated secondary structures (Uzman 2003), although elementary folds (**Figure 3**) are common to the two polymers. An even more fundamental property of RNA is its large repertoire of catalytic activities that makes it far more functional than DNA. The first catalytic RNA was discovered in 1982. RNA enzymes, the so-called ribozymes, have changed our view on molecular biology (Kruger et al. 1982). Right after the function of RNase P was established, more natural ribozymes were discovered *in vivo*. These ribozymes are like ancient fingerprints of the early RNA world and inspired scientists to explore the early form of life (Baer et al. 1988).

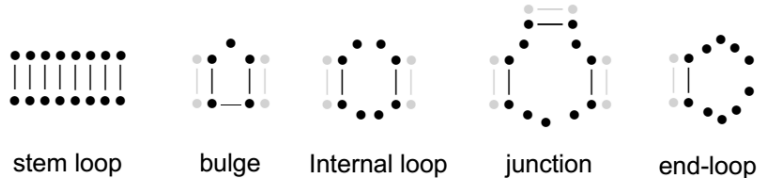



**Figure 3**. Common secondary structure of RNAs. Including: stem, bulge, internal loop, junction and end-loop. Nucleotides are represented by solid balls and hydrogen bonds by a line (Ádám Kun 2015).

The discovery of natural ribozymes has drawn the interest of scientists towards understanding their catalytic mechanisms. Meanwhile, *in vitro* selection was rapidly developing in laboratories and led to the discovery of various artificial ribozymes in past few decades (Chen et al. 2007). It turns out that RNA molecules are far more functional than initially assumed. More and more researchers now believe that RNA molecules are a major player during the evolution of the early life (Pressman et al. 2015). However, some fundamental issues such as the evolution of the genetic code are still a big puzzle. Also, how and why the catalysis of some chemical reactions possibly switched from RNA to protein and how the storage of genetic information switched from RNA to DNA is still debated.

## 2. The RNA world hypothesis

Since Walter coined the "RNA world" 30 years ago, the RNA world hypothesis has become a popular theory. This hypothesis stipulates that there are three stages in the evolution from the RNA world to the modern DNA/RNA and protein world: (**I**) RNA molecules were self-assembled based on their auto-catalytic activities and accumulated by self-replication, recombination and mutation. These processes made them explore new functions. By employing cofactors (ions, peptide, amino acid, etc.) (Johnson-Buck et al. 2011), RNA molecules were able to develop an entire range of enzymatic activities (**Figure 4**); (**II**) RNA adapter molecules recruited from a pool of RNA molecules could bind different amino acids and polymerize them while using RNA templates: protein synthesis began. Co-evolution would promote the RNA/protein world establishment (Copley et al. 2007); (**III**) DNA molecules eventually replaced RNA as the template information holder (Gilbert 1986): the RNA/protein world switched to DNA/protein world.

Because of the irreversibility of evolution, the RNA world has disappeared billions of years ago and it is impossible to find physical traces of this world in the environment of early geological times. However, the RNA world hypothesis is still supported by some existing facts, such as the discovery of natural ribozymes, while the core structure of ribosome which is also based on RNA, and there are plenty RNA viruses (Robertson et al. 2012; Flores et al. 2014). Moreover, *in vitro* ribozyme selection also showed that RNA molecules can catalyze various biochemical reactions.

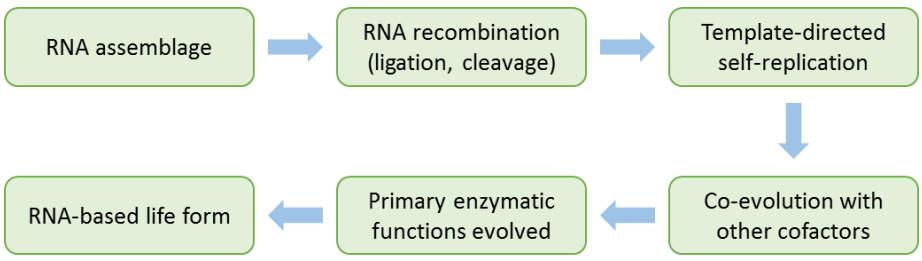



**Figure 4**. The first stage of RNA world evolution.

As described by the hypothesis, there are three critical functions that RNA molecules had to manage during the first stage: 1) self-replication (to undergo natural selection) (Robertson et

al. 2012); 2) catalyzing various chemical reactions (to transform energy) (Morowitz et al. 1988); and 3) aminoacylation (that linked the RNA world to the protein world) (Yarus 2011).

In the very beginning, the RNA evolution started when it gained the ability to replicate, ligate, and/or perform splicing, so that a process equivalent to natural selection could begin (**Figure 5**). Former investigations in the 80s showed that single nucleotide can yield poly-oligomers in the presence of divalent cation at alkaline pH, 0°C after a few days of incubation with a mineral catalyst (Sleeper and Orgel 1979; Orgel 1986). This indicates one way through which RNA could self-assemble. Even though polymerization likely happened extremely slowly and inefficiently, average size nucleotides could accumulate during long geological time scale. Furthermore, molecular cooperation among short RNA molecules would significantly increase the variability of polynucleotide chains through recombination, that might have provided favorable conditions to the evolution of larger macromolecules (Higgs and Lehman 2014). Once the first template-directed replicating ribozyme came about, the growth of polynucleotide chains dramatically accelerated (Ekland and Bartel 1996; Martin et al. 2015). Complementing elongation, self-cleaving is an important way to increase variability during evolution because it allows recombination. Scientists have found many self-cleaving ribozymes either in cells or in viruses that can perform very accurate cleavage. Prominent examples include HDV ribozymes and group I/II intron. They might be ancient ribozymes that survived the transition to the modern biology world (Jimenez et al. 2015).

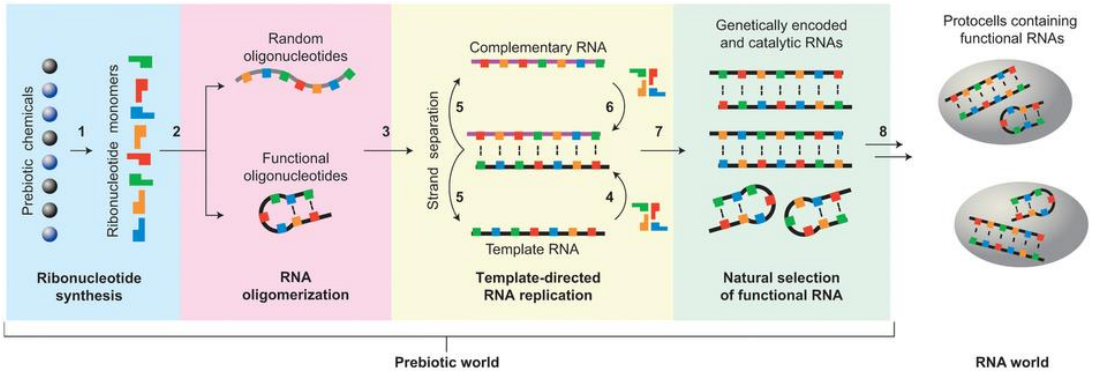



**Figure 5**. (1) Random activated monomers randomly self-assembled into functional and non-functional oligonucleotides (2), which could serve as templates (3) that guide the complementary strand synthesis (4). Inter- and intramolecular recombination allows more copies of RNA transcripts to yield (5) to promote the replication and evolution (6). Being through the natural selection, RNAs could evolved to more sophisticated structures and functions (7), leading to the appearance of the first protocell and the early biocatalysts of the RNA world (8). Illustration from (Hernandez and Piccirilli 2013).

*In vitro* selections achieved since the 90s showed that RNA can accelerate many chemical reactions, such as amide bond formation, aldol reaction, *Claisen* condensation, *Michael* addition, thiol ester formation, *Diels –Alder* reaction and redox reactions (Chen et al. 2007). These functions do not tell how was the early metabolism. However, they indicate that RNA molecules are able to catalyze various reactions in addition to those associated with natural ribozymes (Jäschke and Seelig 2000). Based on many *in vitro* experiments, scientists have proposed chemical models to explain ribozyme catalytic mechanisms, in which the phosphate-ribose skeleton is the scaffold and nucleobases mainly govern ligand binding and construct active core formation (Scott and Klug 1996; DeRose 2002; Lönnberg and Lönnberg 2005). Moreover, these investigations not only helped us understand the function of natural ribozymes, these ribozymes are now also used as tools in cellular research (Cotten and Birnstiel 1989; Ohkawa et al. 1995; Vourekas et al. 2008). Discovering new ribozymes is a goal still pursued by many laboratories today, and will allow us to expand our knowledge about RNA catalysis and the evolution of life from the early RNA world.

When the RNA world reached the gate of stage (**II**), a major transition was about to occur, that led the system to catalyze protein polymerization using RNA as templates. For this transition to occur, some RNA (the early tRNA adaptor) must have had the ability to undergo 3’ aminoacylation, either through their own action (auto-aminoacylation) or by the action of other RNA/peptide cofactors. In the modern RNA aminoacylation system, specific aminoacyl tRNA synthetase catalyze both amino acids activation and tRNA 2’/3’-end transesterification. ATP is employed as the energy driver to allow substrates to pass the reaction barrier (**Figure 6**). While it is still difficult to know how translation was established in the RNA world, it is likely that many of the steps in the process today are at least chemically similar to those that were initially occurring (Ellington et al. 2009). In the absence of aminoacyl tRNA synthetase, it can be conjectured that RNA self-aminoacylation constituted the early solution to the first chemical step of protein synthesis. This question is at the center of the present investigation.

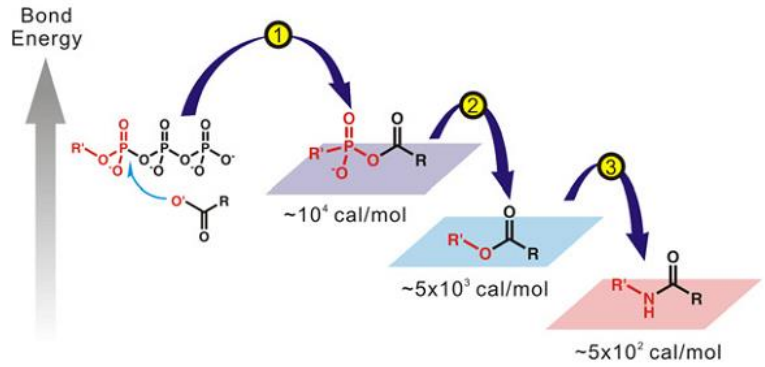



**Figure 6**. Down-hill acyl-transfer reactions in translation. Arrows 1 to 3 indicate the three status of translation: 1) Acyl-activation, 2) hydroxyl group acylation and 3) peptide bond formation. All three reactions can be catalyzed by different ribozymes generated by *in vitro* selection (Schimmel and Kelley 2000; Ellington et al. 2009) in different buffering conditions.

Ribozyme-catalyzed aminoacylation requires three basic elements: amino acids, energy resource and buffering conditions (van der Gulik and Speijer 2015). The availability of free amino acids may have been important already in the early history of life. Mixtures of some simple amino acids, such as glycine, alanine, aspartic and glutamic acid, are easily formed in classical Miller experiments (Miller 1953). These amino acids were likely among the first substrates of ribozyme-directed protein synthesis (Orgel 2003; da Silva 2015). Some of today's amino acids, such as lysine, arginine and tryptophan, are much more complex and were plausibly not present at the origin.

An energy source is required to activate the amino acids and allow them to react with the 3'-end of an RNA. Under normal pressure and temperature conditions, this activation can be fulfilled with the participation of an energy-rich cofactor. Although ATP is the most common of such cofactors in present-day cells, there are some discussions about the presence of ATP in the prebiotic environment (van der Gulik and Speijer 2015), and there is no widely accepted theory about prebiotic activation. The presence of RNA molecules however implies the presence of activated nucleotides. Furthermore, RNA aptamer investigations demonstrated that strong interactions can occur between RNA and ATP molecules (Morii et al. 2002; Vaish et al. 2003; Huang and Szostak 2003; Sazani et al. 2004). Interestingly, most of *in vivo* RNA processing enzymes (RNA ligase, RNA spliceosome, aaRS, etc.) require ATP as cofactor to perform their functions. So far however, no *in vitro* selected ribozyme is capable of "utilizing" ATP as an energy source for activation, and all know self-aminoacylating ribozymes require pre-activated amino acids as substrates. Yet pre-activated amino acids, such as aminoacyl-adenylate, are unstable molecules quickly hydrolyzed in solution (Illangasekare and Yarus

1997). It is thus logical to speculate that early self-aminoacylating ribozymes could also manage the activation step. The quest for an activating – self-aminoacylating ribozyme constitutes a main challenge investigated in the present work.

# Part I *In vitro* selection of ribozymes

## 1. Some aspects of *in vitro* selection and the SELEX methodology

In 1960s, Spiegelman and co-workers observed that the genomic sequence of the RNA bacteriophage Q*β* can dramatically change during a repeated *in vitro* replications by the Q*β* replicase (Mills et al. 1967). These changes are caused by the so-called "serial transfer amplification" during which the RNA of a late experiment is the result of previous RNA amplification reactions. After several rounds of replication, 90% of the RNA turns out to be truncated sequences. This directed evolution experiment is often cited to illustrate how Darwinian evolution could also occur at molecular level during *in vitro* experiments.

This experiment was the premise of the SELEX methodology first explored by Bartel and Szostak in the 90s (Bartel and Szostak 1993). The main purpose of a SELEX procedure is to isolate oligonucleotide sequences (DNA or RNA) that have a special property being repeatedly selected from a large pool of random molecules during selection cycles. A successful design lead to a detectable enrichment of the pools with so-called active molecules, possibly after just a few cycles or up to 15-20 cycles.

In terms of ribozyme *in vitro* selection, the major established methodologies have been classified into three main categories (Breaker 1997).

1) *Distinction by self-modification*. This method was used in many of the early *in vitro* selection experiments. Its goal is to isolate ribozymes based on self-modification. By means of physical separation (through molecular recognition), active variants can be selected depending on their size (like the ligating or cleaving ribozyme, **Table 3**) and/or modification (like the phosphorylation ribozyme, **Table 4**).

2) *Binding to specific targets*. This method has helped researchers to explore numerous RNA aptamers *via* their different affinities to the substrates, which includes proteins, amino acids, ATP and other chemical compounds. These aptamers can also work as probes to target specific substrates and closely-related compounds. They may play the role of inhibitors of metabolic processes or as diagnostic agents (Joyce 1994; Ohkawa et al. 1995).

3) *Altering catalytic properties.* Instead of using pools of random molecules for the selection, some experiments explored the possibility to alter the catalytic function of existing ribozymes. By means of partial randomization or point mutations, researchers were able to engineer the efficiency of natural ribozymes and investigate their catalytic mechanism (Kumar and Ellington 1995).

An *in vitro* selection requires the specification of several parameters defining the SELEX protocol. One of the most important parameter is perhaps the length of the templates. The success of *in vitro* selection experiments critically depends on the design of the starting population of molecules; It requires the initial pools to provide enough molecular and structural diversity for the selection. In that respect, long random sequences can provide more structural possibilities. Long sequences however have a higher propensity to misfolded, and in any case full molecular diversity cannot be reached in test tubes with random track longer than about 28 bases (Pobanz and Lupták 2016). On the other hand, it has been demonstrated that (very) short random track implementing carefully design SELEX protocols may lead to a successful selection (Turk et al. 2010; Pobanz and Lupták 2016). There is thus no clear rule to establish this parameter. In their early successful attempt, Bartel and Szostak did use an extremely large initial RNA pool with 220 random nucleotides present on the template to select ribozymes with ligation activity (Bartel and Szostak 1993). With a well-optimized SELEX protocol, Yarus and co-workers later found a ribozyme specified by as little as five nucleotide that is able to catalyze aminoacylation using only three nucleotides in the active center (Turk et al. 2010). This result, as well as results obtained with self-aminoacylating ribozymes as small as about 25 nt (Illangasekare and Yarus 1999; Lehmann et al. 2007) constituted an argument for us to run a SELEX protocol using pools characterized by small random RNA of the order of 20 to 40 nucleotides.

An issue that is sometimes encountered during SELEX experiments can be mentioned here. It can occur that so-called “selfish RNA” or “mini monsters” emerge and invade the sequence population of random pools subjected to selection. In most cases, selection starts with a pool of RNA with random regions flanked by constant regions that are required for PCR (**Figure 7**). Due to their inherent abundance, some short nucleic acids originating from these constants tracks can invade the selection protocol and replicate more rapidly than others. These small

RNA species may finally dominate the population of amplified molecules (Breaker and Joyce 1994). These false positive signals may overspread the targeted molecules and thus increase the difficulty to isolate active molecules.

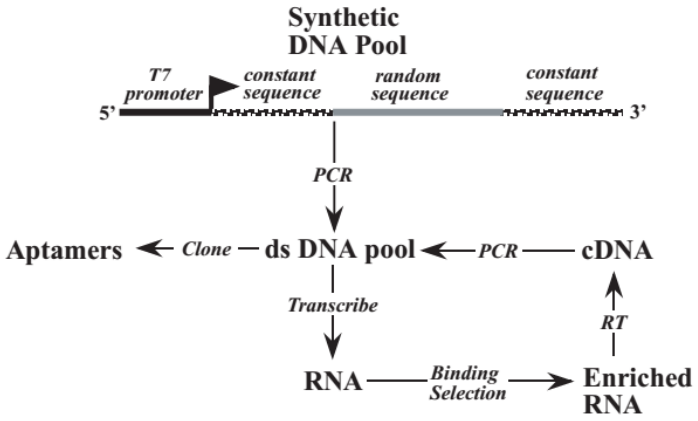



**Figure 7**. Example of typical aptamer pool structure and classical selection process (Wilson and Szostak 1999).

An *in vitro* SELEX experiment is the most powerful strategy so far established to uncover functional nucleic acids, either ribozymes (RNA) or DNAzyme (DNA). It is somehow like a natural selection achieved in a tiny controlled environment, under laboratory conditions. It can be mentioned that *in silico* selection, in which a molecular target can be tested by large numbers of error-and-trial binding tests with different molecules is also a rapidly growing field investigated by bioinformaticians (Könnyű et al. 2015).

## 2. The selection of artificial ribozymes

The diversity of artificial ribozymes found by *in vitro* selection is highly variegated compare with natural ribozymes. In this section, we are going to categorize the artificial ribozymes based on their functions as well as the optimal condition(s) under which they are active. The conditions highlighted in yellow are considered in relation with our protocol to select active ribozymes capable of both amino acid activation (with ATP/GTP) and 3' transesterification.

Since self-amplification would be the beginning of RNA evolution, researchers were quite interested in studying RNA polymerization activity at the early stage of the SELEX development. When Sleeper and co-workers discovered that single nucleotides can aggregate and slowly polymerize (Sleeper and Orgel 1979), the question was: could RNA molecules

rapidly increase in diversity and size, accumulate, and undergo some kind of natural selection? *In vitro* selected ribozymes that can catalyze template-directed elongation are shown in **Table 1**. These ribozymes may represent the ancient semi-conservative replicators.

**Table 1**. Ribozymes with RNA elongation property

| No. | Function | pH | Cofactor | Reference |
|---|---|---|---|---|
| **1** | RNA polymerization | 8.0 | $Mg^{2+}$ | (Ekland and Bartel 1996) |
| **2** | Template-directed polymerization | 8.0 | $Mg^{2+}$ | (Johnston et al. 2001) |
| **3** | RNA polymerization | 8.5 | $Mg^{2+}$ | (Zaher and Unrau 2007) |

Because we aim to isolating auto-aminoacylating ribozymes that must bind ATP and/or GTP as well as amino acid(s), it is appropriate to examine existing single-stranded DNA or RNA molecule that can specifically bind targets such as ATP, peptide or any other organic compound with high affinity. Such oligonucleotides obtained through *in vitro* selection are called "Aptamers". Aptamers are obtained with *in vitro* screening procedures that are similar to those designed for ribozymes. With new high-throughput selection and sequencing methods, aptamer libraries are now increasing exponentially (Dupont et al. 2015; Fraser et al. 2015). Even though aptamers are not catalytic RNA, their binding motifs can help us understand some ribozyme catalytic mechanisms. In relation with our work, we are interested in amino acid aptamers and ATP aptamers (**Table 2**).

**Table 2**. *In vitro* selected aptamers targeted by ATP/AMP and amino acids

| No. | Target | pH | Cofactor | Reference |
|---|---|---|---|---|
| **1** | ATP | 7.6 | $Mg^{2+}$ | (Sassanfar and Szostak 1993) |
| **2** | ATP/CoA | 4.0 | $Mg^{2+}$, $Mn^{2+}$ | (Burke and Hoffman 1998) |
| **3** | cAMP | 7.5 | $Mg^{2+}$, $Mn^{2+}$, $Ca^{2+}$ | (Koizumi and Breaker 2000) |
| **4** | ATP | 7.12 | $Mg^{2+}$ | (Vaish et al. 2003) |
| **5** | ATP | 7.4 | $Mg^{2+}$, NTPs, Spermidine | (Sazani et al. 2004) |
| **6** | ATP | 7.6 | $Mg^{2+}$ | (Huang and Szostak 2003) |
| **7** | GTP | 7.3 | $Mg^{2+}$ | (Carothers et al. 2006) |
| **8** | L-Arginine | 7.6 | $Mg^{2+}$ | (Famulok 1994) |
| **9** | L-Citrulline | 7.6 | $Mg^{2+}$ | (Famulok 1994) |
| **10** | L-Arginine | 7.6 | $Mg^{2+}$ | (Geiger et al. 1996) |
| **11** | L-Tyrosine | 7.4 | $Mg^{2+}$ | (Mannironi et al. 2000) |
| **12** | L-Tryptophan | 7.0 | $Mg^{2+}$, $Ca^{2+}$, glycine | (Majerfeld and Yarus 2005) |

The existence of several kinds of such aptamers suggests that RNA molecules could form binding pockets for amino acid activation with ATP and/or GTP. Known interactions between amino acids and RNA have been reviewed by Yarus and co-workers (Yarus et al. 2009). They may give us hints about the early aminoacylation of RNA and genetic code evolution.

Considering ribozyme families at large, self-processing ribozymes so far constitute the largest family of artificial ribozymes. This family comprises self-cleaving, phosphorylating and aminoacylating ribozymes.

Many *in vivo* selected self-cleaving ribozymes, such as the HDV ribozyme (Riccitelli and Lupták 2013) and the hammerhead ribozyme (Lee et al. 2013) have been engineered by researchers to investigate and improve the efficiency of their catalytic mechanisms. Self-cleaving activity is based on the phosphate-ribose skeleton (Jimenez et al. 2015). An external electron donor attacks the hydrogen atom of a 2' hydroxyl group. The negatively charged 2' oxygen then becomes an active nucleophile that can attack the 3' phosphate and yield a transition intermediate (**Figure 8**). Sharing a similar acid-base catalytic mechanism, ribozymes catalyzing RNA ligation shows another interesting aspect of RNA catalysis (**Table 3**). This key principle of cleaving and ligating is supporting one of the basic catalytic functions of RNA molecule, and is a major reason why RNA molecule is less stable than DNA. RNA-catalyzed ligation and cleavage is the basis for RNA recombination and thus provide opportunities for evolution by increasing molecular diversity.

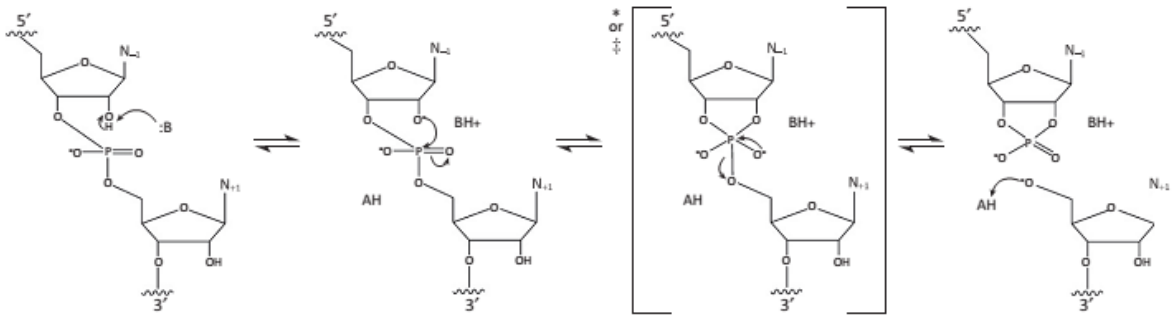



**Figure 8**. A general acid-base mechanism of RNA cleavage. A general base can deprotonate the 2' hydroxyl of the nucleophile, positioned in-line with the 5' *O* leaving group. Then a phosphorane transition intermediate can lead to the transesterification reaction depending on the stability of the transition status. Cleaved upstream RNA will construct a 2'-3' cyclic phosphate and release the downstream as an oxyanion nucleotide. (Jimenez et al. 2015).

**Table 3**. *In vitro* selected ribozymes with ligation and cleavage activity

| No. | Function | pH | Cofactor | Reference |
|---|---|---|---|---|
| **1** | RNA ligation | 7.4 | $Mg^{2+}$ | (Bartel and Szostak 1993) |
| **2** | RNA ligation | 7.4 | $Mg^{2+}$ | (Ekland et al. 1995) |
| **3** | RNA ligation | 7.4 | $Mg^{2+}$ | (Hager and Szostak 1997) |
| **4** | RNA ligation | 7.4 | $Mg^{2+}$ | (Teramoto et al. 2000) |
| **5** | RNA ligation | 4.0 | $Mg^{2+}$ | (Miyamoto et al. 2005) |
| **6** | RNA ligation | 7.5 | $Mg^{2+}$ | (Ohuchi et al. 2008) |
| **7** | RNA ligation | 7.5 | $Mg^{2+}$ | (Fujita et al. 2010) |
| **8** | RNA ligation | 8.0 | $Mg^{2+}$ | (Lie et al. 2016) |
| **9** | RNA cleavage | 7.5 | $Pb^{2+}$ | (Pan and Uhlenbeck 1992) |
| **10** | RNA cleavage | 8.1 | MgOAc | (Williams et al. 1995) |
| **11** | RNA cleavage | 5.0 | no | (Jayasena and Gold 1997) |

Two other main ribozyme families are the phosphorylating and aminoacylating ribozymes (**Table 4**). Phosphorylating ribozymes (Westheimer 1987) may promote enzymatic reactions such as RNA/DNA ligation and RNA degradation *in vivo* by adding phosphate group(s) to RNA targets. Previous *in vitro* selection have identified such ribozymes and shown the phosphorylation for RNA polymerization (Sassanfar and Szostak 1993; Connell and Christian 1993).

**Table 4**. Ribozymes being able to catalyze phosphorylation and aminoacylation

| No. | Function | pH | Cofactor | Reference |
|---|---|---|---|---|
| **1** | RNA phosphorylation | 7.4 | $Mg^{2+}$, $Mn^{2+}$, ATP | (Lorsch and Szostak 1994) |
| **2** | RNA phosphorylation | 7.25 | $Mg^{2+}$, $Ca^{2+}$, ATP | (Curtis and Bartel 2005) |
| **3** | RNA phosphorylation | 6.5 | $Mg^{2+}$, $Ca^{2+}$, $Mn^{2+}$, ATP | (Saran et al. 2005) |
| **4** | RNA phosphorylation | 7.5 | $Mg^{2+}$, $Ca^{2+}$, $Mn^{2+}$, $Cu^{2+}$, ATP | (Biondi et al. 2013) |
| **5** | RNA triphosphorylation | 8.1 | $Mg^{2+}$, trimetaphosphate | (Moretti and Müller 2014) |
| **6** | RNA triphosphorylation | 8.3 | $Mg^{2+}$, trimetaphosphate | (Dolan et al. 2015) |
| **7** | Acyl-transfer | 7.3 | $Mg^{2+}$, Met-*O*-RNA | (Lohse and Szostak 1996) |
| **8** | Acyl-transfer | 7.4 | $Mg^{2+}$, Phe-AMP | (Jenne and Famulok 1998) |
| **9** | RNA aminoacylation | 7.0 | $Mg^{2+}$, $Ca^{2+}$, Phe-AMP | (Illangasekare et al. 1995) |
| **10** | RNA aminoacylation | 8.0 | $Mg^{2+}$, Met-*O*-RNA | (Lee et al. 2000) |
| **11** | RNA aminoacylation | 7.5 | $Mg^{2+}$, Phe-CME/AMP/TE | (Saito et al. 2001) |
| **12** | RNA aminoacylation | 7.25 | $Ca^{2+}$, Phe-GMP | (Illangasekare and Yarus 1999) |
| **13** | Acyl-transfer | 7.0 | $Mg^{2+}$, Phe-p-KK13 | (Xu et al. 2014) |
| **14** | Peptide bond formation | 7.4 | $Mg^{2+}$, Met-AMP | (Zhang and Cech 1998) |
| **15** | Peptidylation | 7.25 | $Mg^{2+}$, $Ca^{2+}$, Phe-AMP | (M. Illangasekare 1999) |
| **16** | Peptidylation | 7.4 | $Mg^{2+}$, Met-AMP | (Sun et al. 2002) |
| **17** | Amide bond formation | 7.4 | $Mg^{2+}$, $Ca^{2+}$, PheCoA | (Li and Huang 2005) |
| **17** | Amino acid activation | 4.0 | $Ca^{2+}$ | (Kumar and Yarus 2001) |

Aminoacylation by ribozymes has drawn a great attention from scientists during last two decades, and is at the center of our investigations. The aminoacyl RNA is the genetic information transporter and an essential tool for protein synthesis: this molecule is central for the evolution from the RNA world to the protein world. Scientists have already uncovered many ribozymes that could have an activity close to the original aminoacyl-RNA synthetase (**Table 4**). Besides, other functions, such as RNA alkylation, thiol ester formation and RNA capping, were also further discovered by *in vitro* experiments, revealing that RNA has a very large potential in catalytic activities (Chen et al. 2007).

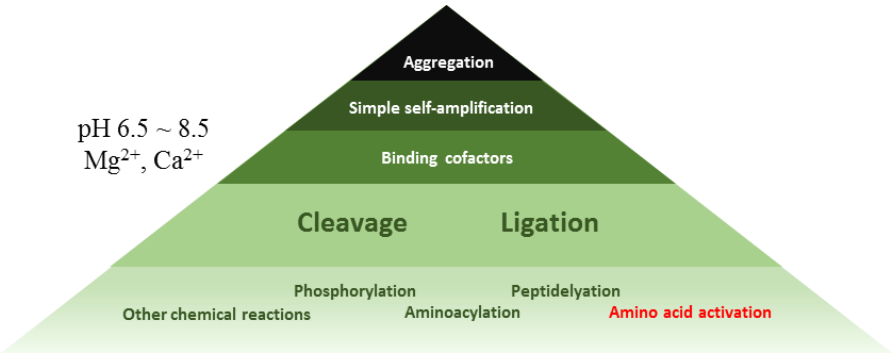



**Figure 8**. Artificial ribozyme identified by SELEX experiments suggest the possibility of an RNA world metabolism. Most of the ribozymes require the $Mg^{2+}$ or $Ca^{2+}$ as cofactors and catalyze reaction at a range of pH value from 6.5 to 8.5. Exceptionally, amino acid activation so far has not been demonstrated at physiological pH (Kumar and Yarus 2001).

Considering a broad image of catalysis in the RNA world (**Figure 8**), most of the ribozymes are active at a pH ranging from 6.5 to 8.5. Catalysis is facilitated by divalent cation. This overview reveals another striking feature: it has not yet been possible to isolate ribozymes catalyzing amino acid activation at physiological pH. In this connection, previously *in vitro* isolated self-aminoacylating ribozymes all require highly concentrated AMP-activated amino acid at the start of the experiments to compensate for its rapid decay at physiological pH (the half-life of Phe-AMP at pH 7 is only a few minutes at 0ºC). How is this issue managed in modern organisms? tRNA aminoacylation *in vivo* is achieved by a series of enzymes called aminoacyl-tRNA synthetase (aaRS). The first step of aminoacylation is amino acid activation with ATP, yielding the activated amino acid (aa-AMP), which remains bound to the enzyme. The aaRS subsequently catalyzes the transfer of the amino acid to the 3'-end of the tRNA (transesterification). In early *in vitro* SELEX experiments, researchers successfully isolated 2'

or 3’ self-aminoacylating ribozymes using synthetic aa-AMP (or with other activating cofactors) as substrate. The question is: where did those activated amino acids come from in the early RNA world environment? Given that transesterification best works at physiological pH, and because aa-AMP is highly unstable at that pH, it can be speculated that, similarly to the aaRS, ribozymes that could successfully manage the activation step should also be capable of catalyzing transesterification. In other words, an aminoacylating ribozyme that could sustainably achieve aminoacylation must be able to manage both chemical steps at the same pH. Uncovering ribozymes that could work as primary aaRS is our main goal. It would fill a huge gap of RNA world hypothesis.

## Part II Identification of aminoacylated RNA from total RNA

### 1. Transfer RNA *in vivo*

As discussed in last section, RNA-catalyzed RNA aminoacylation is the key chemical reaction in the evolution of life from the late RNA world to the modern DNA/protein world. In modern cellular life, tRNA as the major aminoacylated RNA is well known as the key actor of mRNA translation. Besides, tRNA also perform additional functions in gene regulation (Raina and Ibba 2014).

The control and regulation of tRNA aminoacylation is one of the well studied global regulatory system in bacteria. At the translational level, enzymes like RelA in *E. coli* can sense the stalling of protein synthesis that occurs when uncharged tRNA start accumulating due to amino acid starvation. These enzymes activate a downstream gene regulation network to inhibit global gene transcription and activate some other genes related to amino acids synthesis (**Figure 9**, Left) (Ross et al. 2013). Another aminoacylated RNA playing an important role at the translational level is the transfer-messenger RNA (tmRNA), a bacterial RNA with tRNA-like and messenger RNA-like motifs (Komine et al. 1994). The main function of tmRNA is to rescue the translation system by: 1) rescuing the stalled ribosome, 2) degrading the problematic mRNA, and 3) extending the unfinished polypeptide (Janssen and Hayes 2012). This translation quality control system is involved in bacterial development, pathogenesis and stress responses. At the transcriptional level, the level of tRNA aminoacylation can regulate the expression of genes involved in amino acid metabolism through the T-box regulatory system in Gram-positive bacteria (Green et al. 2010). A 5' leader T-box riboswitch is able to bind both the anticodon and the NCCA 3'-end of an uncharged tRNA when this population is high. This results in a read-through of the termination site and the transcription of the full length mRNA (**Figure 9**, Right) (Henkin 2008).

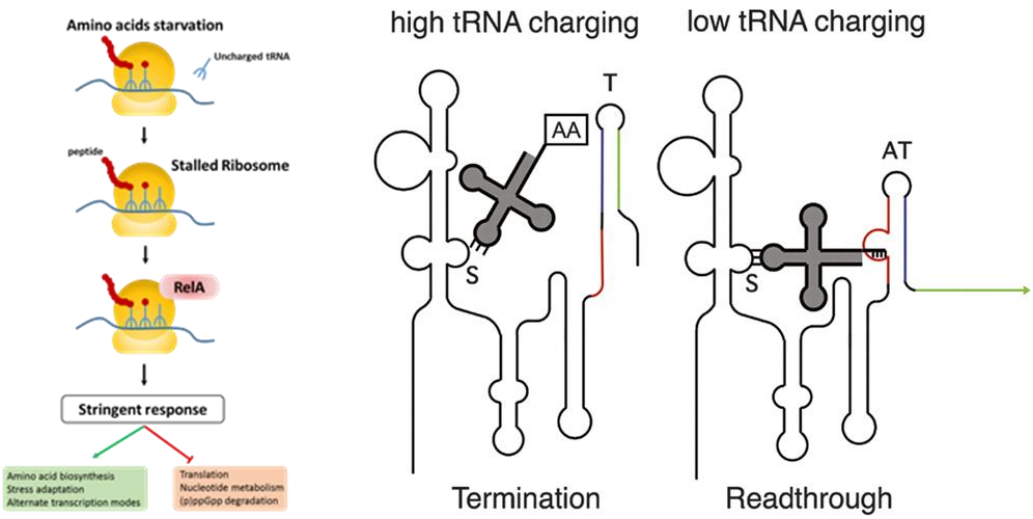



**Figure 9**. Left: Amino acid starvation response in *E. coli* (Ross et al. 2013); Right: The T-box regulatory system (Henkin 2008).

Most of the genes regulated by tRNA are directly involved in translation, and crucial for survival (**Figure 10**). An external stress can cause changes in tRNA gene expression, aminoacylation status, even base modification of specific tRNAs, which in turn regulate other genes at both transcriptional and translational levels (Ibba 2015). It is thus desirable to understand how changes like aminoacylation level and base modifications affect gene regulation.

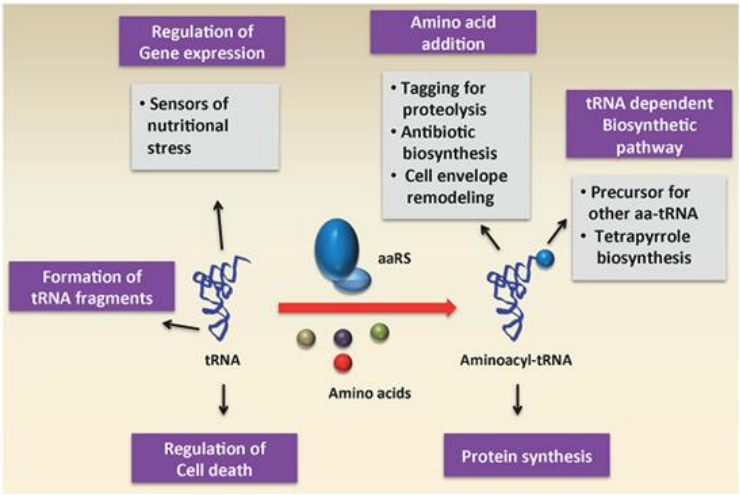



**Figure 10**. Charged and uncharged tRNA plays various roles in living cell, which are mostly link to translation (Raina and Ibba 2014).

Researchers have developed some deep sequencing protocols (Guo et al. 2015; Diebel et al. 2016) to analyze tRNA expression profiles from total RNA. A major difficulty that still

prevents a quantitative and qualitative overview of these profiles is the presence of modified bases on RNA, and tRNA molecules are heavily modified post-transcriptionally (Balke et al. 2015). More than 100 kinds of base modifications have been found so far according to the RNA modification database (http://mods.rna.albany.edu/mods/). These base modifications can increase the molecular stability and regulate the affinity of codon-anticodon interaction (El Yacoubi et al. 2012). Because they can cause stops/pauses of the reverse transcription enzyme, it is not always possible to obtain full length cDNA from these RNA, and that generates biases in RNAseq analysis. Interruption events due to modifications have already been investigated (Motorin et al. 2007). Furthermore, mass spectrometry methods are now available to detect base modifications (Basturea 2013).

Deep sequencing analysis of total RNA have revealed new features related to tRNA degradation: a recently discovered new class of small RNA are the so-called tRNA-derived small RNA fragments (tRFs). These small fragments result from the processing of mature or precursor of tRNAs by RNase P, RNase Z and Dicers , that generates tRNA halves, 5' tRF, 3' CCA tRF, 3' U tRF and 5' leader exon tRFs structures (**Figure 11**) (Lee et al. 2009).

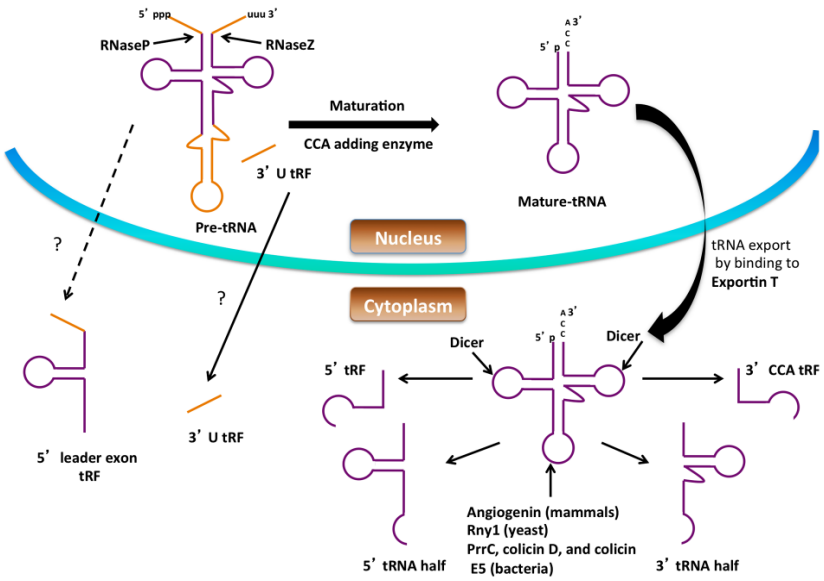



**Figure 11**. tRNA-derived small RNAs (tRFs). Precursor tRNAs need to be processed by RNase P, RNase Z at both 5'- and 3'-ends, as well as CCA-adding enzyme to yield the mature tRNA in the nucleus. Processed fragments from pre-tRNA and mature tRNA is possible to play a role as small RNA in cell. tRNA can further degrades into tRNA halves, 5' tRF, 3' CCA tRF, 3' U tRF and 5' leader exon tRFs, etc. Dash lines and question marks emphasize the unknown mechanisms of formation or transport of these tRFs (Raina and Ibba 2014).

tRFs discovered by high throughput sequencing constitute a class of short RNAs that are second most abundant to microRNAs. These small RNA can be categorized by their size and processing sites. They seem to be involved in many gene regulation and are related to some diseases (Garcia-Silva et al. 2012; Fu et al. 2015). Several facts are already know about tRFs (Kanai 2015): 1) They do not derive from abundant cellular tRNAs, and the numbers of tRFs do not correlate with the parental tRNA gene copy numbers; 2) their fragmentation patterns are depended on their anticodons; 3) the fragmentation patterns can be changed based on environmental conditions; and 4) some tRFs are bound to Argonaute/Piwi proteins (RNA-induced silencing complex). tRFs have been observed in all three domains of life and are expressed by different pathways (Keam and Hutvagner 2015).

## 2. Other possible aminoacylated RNA?

The above overview shows that new features related to tRNA metabolism are still being discovered. Our interest in RNA aminoacylation led us to ask the following question: besides tRNA and tmRNA, are there other types of RNAs that are aminoacylated for functional purposes in modern cells? Some clues suggest this possibility. Shi and co-workers have investigated several truncated tRNA tetraloops that can be recognized and aminoacylated by aaRS *in vitro* (Shi et al. 1992). These tetraloops are derived from tRNA acceptor arm (**Figure 12**), and one may wonder whether some still unidentified tRFs or other RNA could undergo aminoacylation.

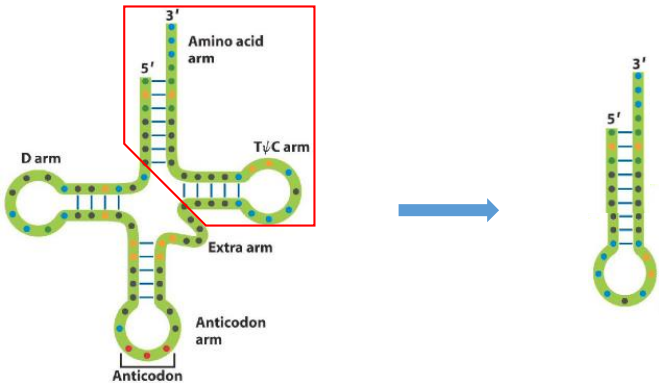



**Figure 12**. tRNA derived tetraloop with minihelix structure that can be recognized and aminoacylated by aaRS.

Very interestingly, some tRNA-like small RNAs have been identified in eukaryotic cells. For example, human and rat U2snRNAs are small RNAs that involved in mRNA splicing, which contain a 3'-end that strongly resembles a tRNA minihelix (the 'top half ' of tRNA) including a CCA ending (Geslain and De Pouplana 2004). Chen and co-workers reported a widespread bacterial noncoding RNA, Y RNAs, that mimics tRNA T-loop and D-loop structures and bear some base modifications similar to that of tRNA. In *Mycobacterium smegmatis*, Y RNA can be processed by RNase P and further modified by CCA adding enzyme (Chen et al. 2014). They suggest that this kind of RNA may mimic tRNA behavior to avoid rapid degradation by RNase *in vivo*. If so, is it possible that this tRNA-like Y RNA could also be aminoacylated? Furthermore, some RNAs found at the 3'-end of plant viruses genome mimic host tRNA structures to support infectivity (Fechter et al. 2001; Hema et al. 2005; Dreher 2009). All these facts suggest that RNA *in vivo* aminoacylation might be more widespread than previously thought. They motivated us to develop an RNAseq protocol specifically designed for the identification of 3' aminoacylated RNA from total RNA. Establishing such a protocol could not only reveal new types of aminoacylated RNA; it will also provide additional information on tRNA regulation and its connection with cell metabolism.

# Chapter I Development of a SELEX method to uncover auto-aminoacylating ribozymes

## 1. Introduction

The SELEX methodology, introduced in the first section, is a technique that allows the screening of pools of random RNA or DNA for molecules with a particular feature. In order to introduce and best justify our protocol of selection of aminoacylated RNA, it is appropriate to present some successful *in vitro* selection experiments that already allowed the isolation of similar ribozymes. Potential restrictions of the method and possible ways to overcome them are also pointed out. An *in vitro* SELEX protocol commonly includes four aspects: initial RNA pool preparation, catalysis processing, *in vitro* selection and re-construct the initial RNA pool by amplification (**Figure 1-1**).

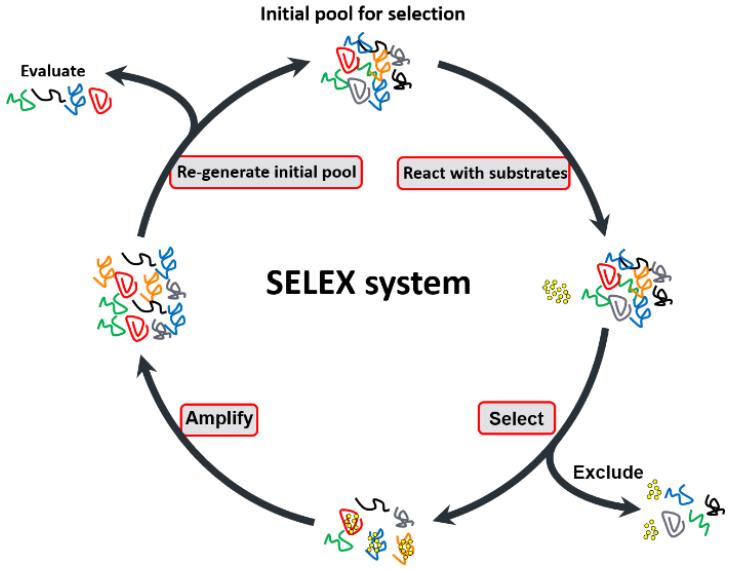



**Figure 1-1**. The common procedure scheme for the *in vitro* selection of catalytic RNAs.

SELEX experiments have allowed at least three research groups (Szostak, Yarus and Suga) to successfully isolate ribozymes involved in the aminoacylation reaction. The two reactions involved in the RNA aminoacylation from free amino acids are the following:

**1)** Amino acid (aa) **activation**: aa + ATP ⇔ aa-AMP + ppi

**2)** 3’ end **transesterification**: aa-AMP + RNA ⇔ aa-RNA + AMP

These two reactions are coupled by the aaRS in the modern translation system, a thermodynamical requirement due to the very low equilibrium constant of the first reaction. Activation is driven forward by the high affinity of the aaRS for aa-AMP and by the coupled second reaction. Contrary to activation, transesterification has indeed a very high equilibrium constant: the free-energy of aa-AMP hydrolysis is about -15 kcal/mol at pH 7.0 and 20ºC (Pascal et al. 2005), resulting in an overall equilibrium constant (reactions 1 and 2) slightly lower than 1 (Berg et al.1961). These thermodynamic considerations show that the difficulty essentially lies in the completion of the activation reaction. Not surprisingly, almost all successful SELEX experiments so far published concern transesterifying ribozymes using pre-activated amino acids as substrates. It is quite illuminating to examine the structure of these identified ribozymes (and related ones) because they provide indications about critical requirements for aminoacylation, and thus for the design of the random pools to implement the SELEX protocol.

**Design of the random pools**

An overview of the published results reveals that the templates designed by different investigators are similar. In general, it includes 50 or 80 random bases flanked by around 20 conserved nucleotides for PCR amplification, which provides around $10^{14}$~$10^{15}$ different sequences in the initial pools (under an appropriate concentration of ~1 $\mu$M). In these former selections, the activity of the isolated ribozymes had to be compatible with these constant tracks, but this issue is often not much discussed. However, these ribozymes have particular 5' and 3' ends that are usually most critical for the activity. Almost all of them have structured bases at the 5'-end and a more flexible, single-stranded 3'-end (**Figure 1-2**) (Illangasekare et al. 1995; Lohse and Szostak 1996; Lee et al. 2000). An initial 95-nt "#29" ribozyme isolated by Illangasekare and Yarus (1997) (**Figure 1-2, C**) could be shrinked to a functional 29-nt ribozyme (**Figure 1-3, A**), revealing that the essential part involved in catalysis is quite small. Furthermore, a study of the 3'-end base-composition of this "29 nt" ribozyme shows that the aminoacylation efficiency can be tuned by changing just a few nucleotides (Lehmann et al. 2007). On the 5' side, it was shown that a terminal triphosphate (always present with T7 transcripts) is essential to the activity of this ribozyme family (**Figure 1-3, A**). Altogether, these

observations suggest that it is crucial to minimize the (arbitrary) constant tracks at both 3' end 5' ends in the design of the random pools.

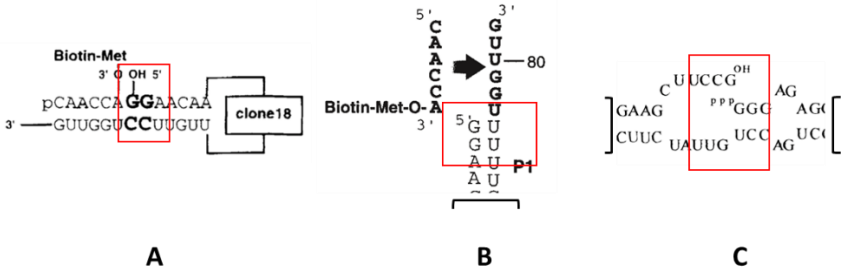



**Figure 1-2**. The 3' and 5'-ends of *in vitro* selected ribozymes from Szostak (**A**), Suga (**B**) and Yarus (**C**) teams. The red square circle indicates where esterification (**A, B**) and aminoacylation (**C**) occurs. All the 5'-ends are base-paired and their sequences are part of in the constant region in the initial pools. The 3'-ends are mostly free. "[" indicates the abbreviated parts. **A**, **B** and **C** are adopted from (Illangasekare et al. 1995; Lohse and Szostak 1996; Lee et al. 2000).

The above conclusions most likely triggered Yarus and co-worker to try an improved SELEX protocol, with which they could identify much smaller self-aminoacylating ribozymes from initial pools characterized by entirely random 3' ends. Working with this protocol, they found a series of ribozymes that catalyze aminoacylation with only three conserved nucleotides (**Figure 1-3, B**) (Chumachenko et al. 2009). The identified ribozymes had an 8-base junction domain (**Figure 1-3, B,** Red rectangle) present in the middle of the sequence. These candidates could be simplified into a 9-nt "GUGGC/GCCU" base paired ribozyme that can achieve aminoacyl-, peptidyl-, dipeptidyl- reactions when highly concentrated Phe-AMP is present (**Figure 1-3, C**) (Turk et al. 2011). Similar to the previously selected ribozymes, the 5'-end is structured. The truncated C3 ribozyme (**Figure 1-3, C**) however has an unstructured 5'-end. The structural stability of the active site is achieved by setting a low temperature (4ºC) during incubation.

These former investigations provided unevaluable information that guided us in the design of our random pools. Prominent considerations are the following:

1) Although full randomization is desirable, the 5'-end sequence of the pools may not require to be strictly randomized. T7 transcript must all start with $^{5'}$GG for an efficient transcription, a constraint that seems to be an advantage in terms of secondary structure stability.

2) The 3'-end sequence of the pool is critical for transesterification: maximizing randomization may allow our SELEX experiment to have access to a higher diversity of

ribozymes. This is especially important since we seek to isolate ribozymes with dual catalytic activity (activation and transesterification).

3) The size of the random track may not need to be extremely large. Small motifs can be sufficient for catalysis (**Figure 1-3**).

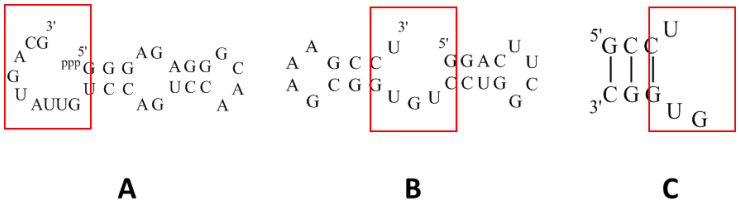



**Figure 1-3**. Small ribozymes derived from *in vitro* selection candidates. **A**: Yarus 29-nt ribozyme derived from "Isolate 29 RNA" (Illangasekare and Yarus 1997). **B**: C3 ribozyme selected by Yarus group in 2009 with three conserved bases (Chumachenko et al. 2009). **C**: further modified a mini ribozyme derived from C3 ribozyme (Turk et al. 2011).

**Incubation conditions and activating cofactor(s)**

The choice of particular incubation conditions set to promote ribozyme aminoacylation may be guided by the examination of the chemistry being involved, and by previous studies on such ribozymes.

Transesterification is based on a classic "acid-base" chemistry mechanism (Hatano and Ishihara 2013). A Lewis acid coordinates with the oxygen of a carboxyl group to reduce the electron density of the carbon atom. The 2' or 3' oxygen atom of an RNA attacks the carbonyl group, and the leaving group is subsequently released (**Figure 1-4, A**). Ribozyme-catalyzed aminoacylation is also based on this mechanism. The Yarus group has proposed a possible catalytic mechanism for the C3 ribozyme based on computed energy minimization with solvent (**Figure 1-4, B**), where the G14 nucleotide plays the role of the M-B reagent (**Figure 1-4, A**) to catalyze the U26 attack on Phe-AMP and the release of AMP.

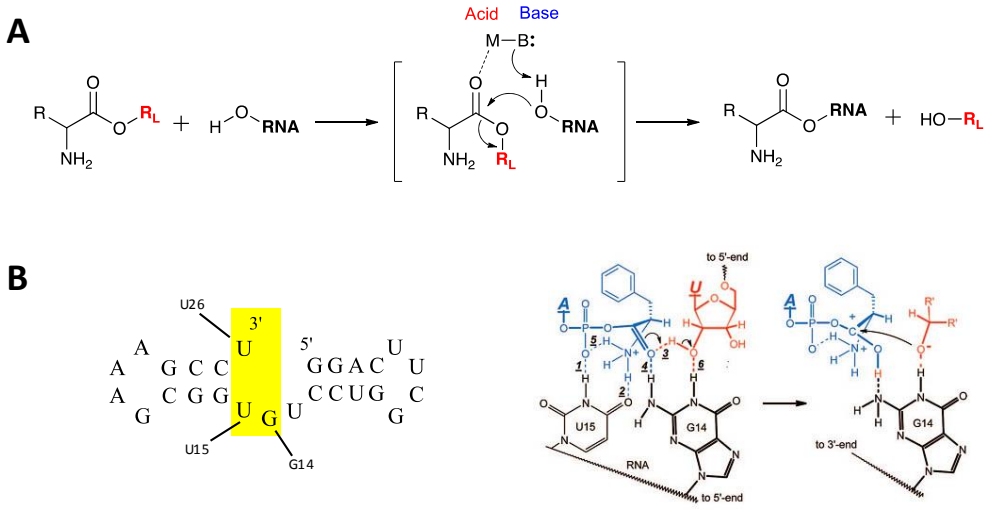



**Figure 1-4**. **A**: Acid-base mechanism of RNA aminoacylation. $R_L$ represents the leaving group. "M" Lewis acids. "B:" organic/inorganic bases or electron donors. **B**: C3 ribozyme and proposed catalysis mechanism.

At least three aspects have to be considered for aminoacylation: 1) the presence of a Lewis acid, 2) the nature of the activating cofactor and 3) setting an optimal pH.

Lewis acid usually refers to cations like $Na^+$, $K^+$, $Mg^{2+}$ and $Ca^{2+}$. Are these cations always required for aminoacylation activity? In an early study on a 5'-self capping ribozyme, the Yarus group discovered that a selected "isolate 6 RNA" is a Calcium dependent ribozyme while the presence of Magnesium even inhibits catalysis (Huang and Yarus 1997). On the other hand, the later investigated tiny ribozyme "GUGGC/GCCU" (**Figure 1-3, C**) does not require divalent cations at all; the presence of Magnesium only slightly increases the reaction rate (Turk et al. 2011). Nevertheless, some ribozymes such as the Yarus "#29 isolate" (**Figure 1-3, B**) requires both Calcium and Magnesium for activity, while other isolates show different divalent-dependence profiles (Illangasekare et al. 1995). In addition of being potential Lewis acids, divalent cation helps RNA molecule to properly fold by neutralizing the negative charge of polyanionic backbone and support out-sphere binding to stabilize the secondary structure (Saito and Suga 2002; DeRose 2003; Sigel and Pyle 2007). From these results, a conservative strategy is to include both Magnesium and Calcium in the incubation buffers in addition of monovalent cations.

Because a carboxyl group of an amino acid may not react with a hydroxyl group under standard conditions due to a very high energetic barrier (Zhang and Cech 1997; Li and Huang 2005; Murakami et al. 2002; Niwa et al. 2009; Morimoto et al. 2011), former studies on

aminoacylating ribozymes all used pre-activated amino acids to allow transesterification. The commonly employed leaving group, such as AMP, CoA (Coenzyme A), CME (Cyanomethyl ester), CBT (*p*-chlorobenzyl thioester), DBE (*3,5*-dinitrobenzyl ester), etc., are all produced by organic synthesis. All these activated forms may easily react, but some of them (such as AMP-aa) have very short lifetimes and are quickly hydrolyzed in solution. In terms of efficiency, a system implemented with ribozymes requiring pre-activated amino acids as substrates would be extremely inefficient and unlikely to occur in a primitive environment. It is thus reasonable to conceive the possibility that some still unknown ribozymes could manage both activation and aminoacylation.

Choosing an appropriate pH is another critical issue. The aminoacylation reaction requires physiological pH for optimal activity; a higher pH will reduce the half-life of active acyl bond (like Phe-AMP) and also increase the instability of the aminoacyl-RNA ester bond. A low pH will reduce nucleophilic activity of 2'-/3'-hydroxyl groups (Murray et al. 2002), a phenomenon clearly established in the case of ribozymes isolated by Yarus and coworkers (Illangasekare and Yarus 1997).

**Is low pH a requirement for amino acid activation?**

The ribozyme-catalyzed activation of amino acids remained an unsolved problem for many years since the discovery of the first aminoacylating ribozymes in 1995. In 2001, Yarus and coworkers came up with another breakthrough in the ribozyme field with the identification of an RNA (called KK13; **Figure 1-5, A**) able to catalyze amino acid activiation by using the triphosphate group at the 5'-end of the RNA. The pH was set to 4.0 during selection, and the isolated ribozyme, that requires Calcium, was also found to best work at that pH (Kumar and Yarus 2001). This is so far the only known RNA that can catalyze carboxyl group activation.

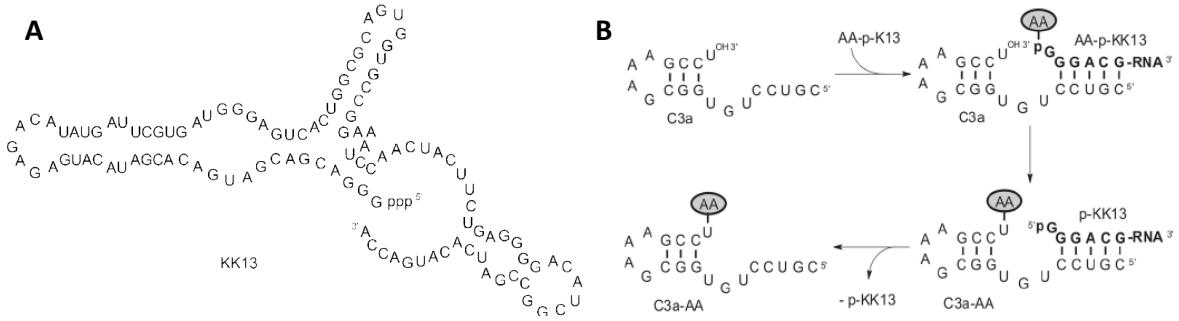



**Figure 1-5**. **A**: Mfold predicted secondary structure of KK13. **B**: Experimental design of two cooperating ribozymes that process amino acid activation and aminoacylation (Xu et al. 2014).

Because this ribozyme functions at pH 4.0 while aminoacylating ribozymes are most efficient at pH 7.0, a combination of both reactions could not be envisaged in a same aqueous media. After an unsuccessful attempt to isolate RNA with both catalytic activities at the same pH, Xu and co-workers established a smart experimental design to combine two ribozymes and two reaction conditions to successfully achieve amino acid activation and aminoacylation (Xu et al. 2014). In this strategy, an amino acid is first activated with the KK13 ribozyme. They then re-designed the C3 ribozyme to allow base pairing with the 5' end of the KK13 ribozyme, that yielded a truncated C3a (**Figure 1-5, B**). The C3a variant enables the activated amino acid to be close to the U26:U15G14 catalysis core of C3a, and thus promote transesterification. This work is meaningful because this is the first time that researchers successfully produce aminoacylated RNA by only using RNA as catalytic cofactor.

However, these experiments can only occur sequentially. The KK13 ribozyme is only functional at pH=4 while transesterification happens at physiological pH range. The reason for which a pH of 4.0 was chosen in the selection that yielded KK13 is that acetate groups cannot react with a triphosphate at physiological pH if a positive charge is not around for coordination (**Figure 1-6, A**). In modern cells, the aaRS catalyze amino acid activation with ATP through the use of magnesium and arginine or lysine residues to manage the negative charge of the triphosphate (**Figure 1-6, B**). It means even if the aaRS work at apparent pH around 7, there is an electron-deficient environment inside the catalytic center (or Lewis acidic environment). Likewise, Yarus KK13 ribozyme works at mildly acidic pH (i.e. in an $H^+$ -rich environment) with 5 mM calcium. Possibly, a calcium-phosphate interaction can neutralize the negative charge of $\beta$- and $\gamma$-phosphate. Moreover, 80% of the $\alpha$-phosphate group is protonated at pH 4 (Corfù and Sigel 1991), which further decreases the electronic repulsion between acetate and triphosphate groups (**Figure 1-6, C**).

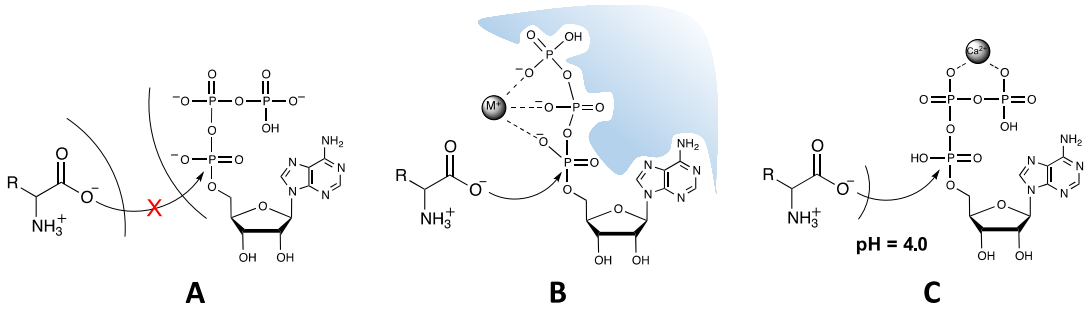



**Figure 1-6**. Amino acid activation by ATP. **A**: At physiological pH, both ATP and an aminoacyl group are negatively charged molecules. Nucleophilic attack is impossible. **B**: aaRS enzymes employ the magnesium as well as positively charged amino acids side chains to bend and significantly reduce the negative charge density of triphosphate to facilitate activation. **C**: Yarus KK13 ribozyme requires calcium coordination pH=4.0 for optimal activity. Under these conditions, the $\alpha$-phosphate has been protonated (Corfù and Sigel 1991). It is possible for aminoacyl group attacking. "(" indicates the repulsive potential surface. Blue color represents a partial enzymatic environment.

A fundamental problem stands here. If we provide enough positive charges during *in vitro* selection, is it possible to achieve amino acid activation at physiological pH and thus work in conditions that are optimal for aminoacylation? Seeking for a ribozyme that can truly functions as a primary aaRS is the significant challenge investigated in our work.

# 2. Experimental design

## 2.1 Overview of the protocol

The protocol is divided into two parts. The first part consists of the incubation of the random RNA pool in a solution comprising a mixture of 19 amino acids, activating cofactors (ATP and GTP) and various monovalent and divalent cations. This incubation is designed to promote self-aminoacylation on active RNA molecules of the pool. In a second part, a procedure of oxidation and deacylation followed by adaptor ligation at the 3'-end and RT-PCR is performed to selectively amplify aminoacylated RNAs.

**Figure 2-1** outlines the individual steps constituting these two parts:

1. The SELEX experiment starts with the synthesis of a pool of RNA *in vitro* transcribed from a series of cDNA templates with 10 to 30 random positions flanked by very short constant tracks.
2. After incubation in a buffer with the amino acids, ATP/GTP, different ions and polymers, the mixture of RNA is collected for selection.
3. The selection step has two parts: First, RNA is incubated in a solution of sodium periodate to oxidize the 3'-end of molecules with no aminoacyl ester attached to them. Periodate oxidation as a popular strategy has been used in molecular biology to oxidize the *cis*-diols groups at the 3'-end of RNA. It is also applied in many RNA *in vitro* experiments to distinguish the aminoacyl and non-aminoacyl tRNA, introduce external modified bases into RNA, or to detect post-transcriptional modifications (Zaborske et al. 2009; Murakami et al. 2002; Kurata 2003; Kawano et al. 2012). Any RNA with free 3'-*cis*-diol group of the ribose is oxidized into di-aldehyde groups, while an RNA aminoacylated at the 3'-end is protected from oxidation. Subsequently, deacylation releases the amino acids from these RNA, making the diols at the 3'-end available to adaptor ligation during the next step.
4. RNA with free non-oxidized 3'-end ligate with a 5'-phosphorylated RNA adapter carrying a specific restriction site near the 5'-end junction. (Adapters design see **Section 2.5**)
5. Reverse transcription generates a cDNA that can readily undergo another ligation step.
6. A double-stranded DNA anchor carrying the T7 promoter sequence ligates with the cDNA

using a few overhanged bases for base paring. (This step is further optimized in the final protocol. See details in **Section 2.6**)

7. The fully structured cDNA including a T7 promoter and an adapter sequence with selected target in between is ready for PCR amplification.
8. The treatment with a restriction enzyme removes the adapter part. A new pool of RNA is transcribed from these templates and allows to start a next selection cycle.

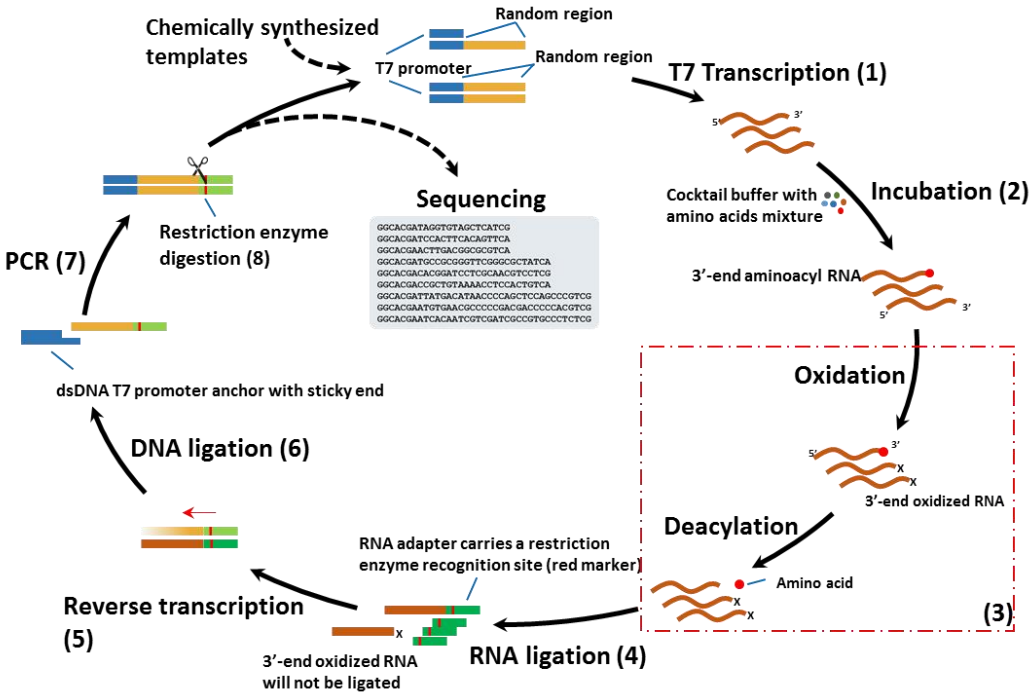



**Figure 2-1**. SELEX procedure: 1) Chemically synthesized DNA templates or treated PCR products which include very short conserved bases are used to transcribed into a pool of random RNA; 2) Pool of RNA incubates with complex buffer; 3) After incubation, RNA is treated by periodate. The 3'-end non-aminoacylated RNA is oxidized. Followed by deacylation, the amino acid is released from the 3'-end of aminoacyl RNA and make the 3'-end available for adapter ligation; 4) 3'-end non-oxidized RNA ligates with a carefully designed RNA adapter carrying a restriction enzyme recognition site by T4 RNA ligase; 5-7) Reverse transcribed DNA undergoes T4 DNA ligation with a dsDNA T7 promoter anchor to construct the full length of cDNA and ready for PCR amplification; 8) Restriction enzyme treatment removes the adapter part and generates the new DNA templates for next round selection.

## 2.2 Template design

In the process of establishing a SELEX procedure to isolate active RNA molecules, the length of the random track and that of the conserved regions of the RNA templates constitute a critical issue.

The size of the random track in most *in vitro* SELEX experiments varies from 20 to 200

nucleotides (Ruckman et al. 1998; Lee et al. 2000; Kumar and Yarus 2001; Thiel et al. 2011). In theory, RNA pools with more random positions provide a higher diversity and more possibilities to the selection, but this strategy results in less copies of each unique sequence present among the whole population at a given molar concentration (Yarus 2005). The issue of how the length of initial pool influent the *in vitro* selection has been recently reviewed by Pobanz and co-workers (Pobanz and Lupták 2016). Another aspect to integrate is the efficiency of the T7 RNA polymerase. This enzyme is more efficient in the transcription of long sequences compared to shorter ones, the lower limit being around 20 nucleotides. Long transcripts provide more structurally diverse binding sites but they may more easily miss-fold during the renaturation step in each selection cycle. Short transcripts can possibly provide active molecules, but set limits to the possible structures of an active site. It has been suggested that autocatalytic RNA networks are more likely to begin from monomers accumulation and short substrate interaction (Gilbert 1986; Yarus 2005; Higgs and Lehman 2014). Considering all these aspects, we decided to implement a selection protocol with 4 pools of RNA molecules with a size of 19, 23, 30 and 37 nucleotides while maximizing the relative portion of the random tracks.

### 2.2.1 Initial design

Our initial design was rather simple. In order to minimize the burden of the constant tracks at both 5' and 3' ends of the molecules, we choose to work with almost completely random RNA sequences, except for two Gs at the 5'-end, required for T7 transcription (**Table 2-1**). Studies on a 3'-end self-aminoacylating ribozyme have indeed shown that the length and the base composition of the 3'-end extension have a significant effect on their activities (Illangasekare et al. 1995; Lehmann et al. 2007). Furthermore, the composition of the 5'-end of these RNAs, including the presence of a 5'-triphosphate, may also dramatically impair the active site of this type of ribozyme (Illangasekare and Yarus 1997).

One major difficulty in the protocol related to the randomness of the RNA sequences is the specificity of reverse transcription. Random region can cause miss priming during the annealing step of reverse transcription. The final PCR could easily yield primer dimers as by-products. We had to test different RNA adapters to solve this issue (**Section 2.5**).

Another main problem associated with these random libraries is the difficulty to keep the size of the pools constant. Because the T7 RNA polymerase has non-template extension activity and may drop off the template before the end of transcription, the size of the RNA pool will turn highly heterogeneous already after one or two cycles (Chumachenko et al. 2009). In the worst cases, up to 50% of the *in vitro* transcribed RNA may contain so-called *n*+1 and *n*+2 nt sequences (Milligan et al. 1987; Krupp 1988). T7 RNA polymerase can also bypass up to ~20 nucleotides gaps before resuming transcription (Zhou et al. 1995). These events are clearly visible after selection during gel analysis, where smears are observed instead of clear bands.

**Table 2-1**. Initial design of the RNA pool

| Name | RNA Sequence (5'→3') |
|---|---|
| **N16**[a] | GGN NNN NNN NNN NNN NNN |
| **...** | **...** |
| **N45** | GGN NNN NNN NNN ..... NNN NNN NNN NNN NNN |

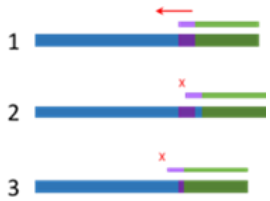
[a] different number indicates the length of the template

Another possibility to generate 3'-end homogeneous transcripts is to add the sequence of the HDV self-cleaving ribozyme on the DNA template after the 3'-end. In that way, a clean and unique 3'-end can be generated after transcription. It however requires a PCR with extended primer PCR and extra gel purification procedures (Schürer et al. 2002; Chumachenko et al. 2009). Because our RNA pools are very short, these extra procedures would imply a significant loss of material during each purification step. In order to prevent transcripts with alternate sizes from invading the selection process, the terminal nucleotides of the sequences were set to very short conserved bases, either SCA or YCCA (**Table 2-2**). During reverse transcription, by using a primer of the adapter with an overhang complementary to these nucleotides, most of these alternate transcripts could not undergo reverse transcription (**Table 2-2**, figure).

**Table 2-2**. Template designed with conserved bases at the 3'-end

| RNA Sequence (5'→3') |
|---|
| GGN NNN NNN NNN ..... NNN NNN NNN S[a]CA |
| GGN NNN NNN NNN ..... NNN NNN NNY[b] CCA |

[a] S=C or G; [b] Y=C or U;




**Right figure.** Template selectivity achieve by the overhang or reverse primers: 1. When template with the correct size is ligated to the adapter, reverse transcription and/or PCR can occur. 2 and 3: when aborted or extended transcripts are ligated to the adapter, mismatch(es) with the overhang of a reverse primer may stop either reverse transcription or PCR. Blue represent selected RNA oligo; Dark purple indicates the conserved bases at the 3'-end; Dark green represents the RNA adapter; The upper strand is the reverse primer (light green = primer; purple = overhang); Red arrow shows the direction of reverse transcription and the red cross symbolizes RT or PCR interruption.

Another major difficulty related with the amplification of essentially random RNA sequences is the requirement of an adaptor ligation at the 5'-end of the transcripts. In order to prevent side-products of RNA ligation from occurring, this second ligation was attempted on cDNA after reverse transcription. Only two C residues could specify the 3'-end of these cDNAs. We attempted to ligate them with a DNA anchor (**Table 2-3**, orange symbol) containing a 3'-end overhang of two Gs constituting sticky end. Despite all our attempts in various conditions, we could not succeed in obtaining a final construct that could undergo a final PCR. It turns out that the yield of this final ligation step was always too low, and that other ligation events would always prevent us from obtaining the desired PCR product. Especially, the DNA anchor would ligate to the excess of primer of the adapter (**Table 2-3**) and generate primer-dimer as the majority PCR product. Probably, only two G residues 3' overhang could easily cause unspecific ligation between the DNA anchor and the primer of the adapter even if they were designed not to be compatible with each other. After months of unsuccessful attempts, we decided to abandon the idea of RNA transcripts with almost totally random 5'-end.

We re-designed the 5'-end while designing more constant positions in order to promote specific DNA ligation. We assumed that a DNA anchor mainly constituted by purine residues would conflict with the purine region of the primer of the adapter and prevent unspecific DNA ligation from happening between these two species.

**Table 2-3**. Templates carrying more constant positions at the 5'-end

| RNA Sequence (5'→3') |
|---|
| GGM[a] ANN NNN NNN ..... NNN NNN NNY CCA |
| GGC ACN NNN NNN ..... NNN NNN NNY[b] CCA |


[a] M=A or C; [b] Y=C or U;

**Right figure**. DNA ligation step: 1. Expected cDNA (3'→5') ligation with the DNA anchor. 2. Unspecific ligation between the DNA anchor and primer of the adapter. Orange symbol represents the double-stranded DNA anchor with a sticky end of overhang compatible with 3'-end of cDNA, the overhang could be either "GGMA" or "GGCAC" (5'→3') for instance; Light blue is the cDNA sequence transcribed from selected RNA target; Light purple and light green is the primer of the adapter (3'→5'), in this case the purple part represents the sequence "RGGT".

A control during which the whole procedure was tested without template (i.e. only with the primer of the adaptor and the DNA anchor) generated a strong by-product on the gel (**Figure 2-2**) with a length corresponding to a primer-dimer. After several unsuccessful attempts, we finally abandoned the strategy of applying two ligations (see **Section 2.6**).

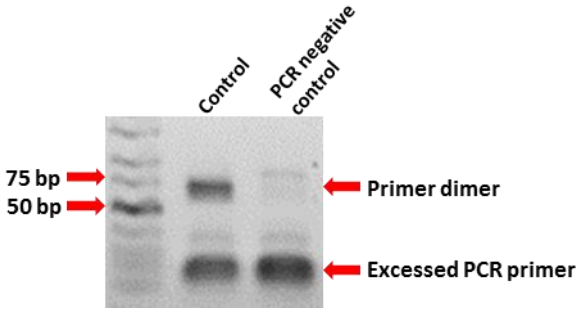



**Figure 2-2.** By-products test: the left lane shows the primer dimer yield by a control which had processed the whole procedures with only the primer of the adapter and DNA anchor; the right lane is a PCR negative control with only two PCR primers.

### 2.2.2 Final design

In order to successfully achieve the final PCR amplification, we finally opted for a solution with a small stretch of constant positions at the 5'-end of the sequences of the pools. This constant track would allow the overhang of a forward primer to prime directly on the cDNA without the need of a second ligation. Although the arbitrary choice of these constant positions could be critical to the success of our SELEX protocol, we could not conceive any other feasible option. We tested other methods to add the T7 adapter to the 5'-end of RNA template by modifying the overhang of DNA anchor, one step cDNA elongation and poly-A tailing, and they all failed. A new survey of the literature made us later realize that other groups were likely already

confronted with this issue. In a recent SELEX investigation whose aim was to isolate new self-aminoacylating ribozymes, Yarus and coworkers also opted for a constant track at the 5'-end, while the 3'-end was completely random (Chumachenko et al. 2009). They however did not try to minimize the size of the 5' track, and kept as much as 20 constant positions.

**Table 2-4**. Template design with 10 conserved nucleotides

| RNA Sequence (5'→3') |
|---|
| **GGC ACG A**NN NNN ..... NNN NNN NNN **UCR**[a] |
| **GGC GAC G**NN NNN ..... NNN NNN NNN **UCR**[a] |

[a] R=A or G

Our carefully re-designed 5'-end contains 7 conserved bases (**Table 2-4**) that turned out to be long enough to ensure a correct priming during PCR with the overhang of a T7 primer. This "***Short primer PCR***" method to avoid primer dimer aggregation is introduced in detail in **Section 2.7**. Shorter constant tracks were tested, and could not allow the PCR amplification of our construct.

A final refinement in the design of our pools concerned the choice of the terminal 3' residues. Our initial 3'-end design was "UCR", where R (purine) stands for A or G (**Table 2-4**). The rationale for a purine at the 3'-end is that they may bind RNA more strongly than a pyrimidine, and thus could better stabilize an active conformation of the 3'-end nearby a catalytic site. Also, the 3'-ends of modern tRNA bear a terminal adenosine, which may be an indication of such requirement.

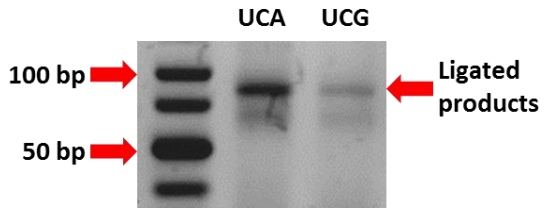



**Figure 2-3**. RNA ligation between the RNA template and RNA adapter. RNA template which is ending by "UCA" (left) shows higher efficiency than "UCG".

By using different overhang primers of the adapter during our test experiments, we observed a much stronger amplification of "UCA" compared with "UCG" in the final PCR,

possibly because the "UCA" ending is more compatible at the RNA ligation level (**Figure 2-3**). For this reason, the UCA ending was selected.

**Table 2-5**. Final template design

| Name | Sequence | *n* (nmol) | Diversity | Copies per species |
|---|---|---|---|---|
| **N19** | GGCACGANNNNNNNNNUCA | 0.1 | $2.6\times10^{5}$ | $2\times10^{8}$ |
| **N23** | GGCACGANNNNNNNNNNNNNNUCA | 0.1 | $6.7\times10^{7}$ | $1\times10^{6}$ |
| **N30** | GGCACGANNNNNNNNNNNNNNNNNNNNNUCA | 0.1 | $1.1\times10^{12}$ | 60 |
| **N37** | GGCACGANNNNNNNNNNNNNNNNNNNNNNNNNNNNUCA | 0.1 | $1.8\times10^{16}$ | 1 |

The final design of our pools is shown in **Table 2-5**. To establish the total length of the initial pools of RNA, more than 50 different cDNA templates were preliminarily tested (data not show). Since the T7 polymerase has a rather low yield with DNA templates shorter than 20 nt, we chose 19 nt as the minimum size for one of these pools. Meanwhile, our "***Short primer PCR***" strategy is not suitable for amplifying cDNAs longer than 70 nt. Furthermore, it is reasonable to keep the length below 50 nt. When we prepare the initial pool of random RNA, the optimal amount of initial cDNA template we used for T7 transcription (0.1 *n*mol) includes $6.02\times10^{13}$ sequences, corresponding to the total diversity obtained with 23 random positions ($4^{23} = 7\times10^{13}$).

## 2.3 Composition of the incubation buffers for aminoacylation

Cofactors, such as ions, amino acids, peptides, ATP/GTP, nucleic acids, even water, are highly essential for ribozyme's structural folding and chemical catalysis. Natural ribozymes, such as Group II intron, RNase P and HDV ribozyme, all require an appropriate pH value as well as a certain amount of $Mg^{2+}$ to be active. Group I intron additionally requires GTP as cofactor in the first step of the splicing reaction. In our experimental design, several cofactors were added as components of the cocktail buffer to stimulate aminoacylation catalysis.

Since the "*Primordial soup*" theory was firstly proposed in 1924, the original life environment has been investigated in laboratories (Bailey 1938b). The negatively charged RNA chain requires the presence of various cations for proper folding and catalytic activity. In

cooperation with different functional groups such as carbonyl, amino, hydroxyl, sulfanyl and phosphate, they may assist align the nucleophile with the reactants by reducing entropic barriers (Sigel and Pyle 2007; Schnabl and Sigel 2010).

### 2.3.1 Ions

Monovalent cations such as $Li^+$, $Na^+$, $K^+$ and $NH_4^+$, are the most abundant ions in both the organisms and the natural environment. Monovalent cations are not only required for the ribozymes to adopt a proper folding (Woodson 2005; Jiang et al. 2006) but also possibly responsible for their catalytic functions (Murray et al. 1998; Hanna and Doudna 2000). A recent research investigating crystal structures of the hammerhead ribozyme derived from *Schistosoma mansoni* has demonstrated that monovalent cations such as $Na^+$ could directly and specifically replace a divalent cation in the active site (Anderson et al. 2013). These coordinated monovalent ions could contribute to the ribozyme catalysis as a Lewis acid to promote deprotonation of OH group or to neutralize the negative charge on the phosphoryl oxygen of the transition state (Ke et al. 2007). Characterized by low atomic radius and charge density, monovalent cations are however far less efficient than the divalent ones; the HDV ribozyme requires up to 2-5 M of monovalent salts instead of $Mg^{2+}$ to support efficient cleavage (Perrotta and Been 2006). Obviously, such a high concentration is hardly achieved in natural environment. Thus certain divalent cations have to be introduced into the selection to ensure the activity of the ribozymes.

The monovalent composition of our incubation buffer is 100 mM $Na^+$ and 100 mM $K^+$, which is close to the concentration range of usual enzymatic reactions and organism cellular living environment (Linzell and Peaker 1971).

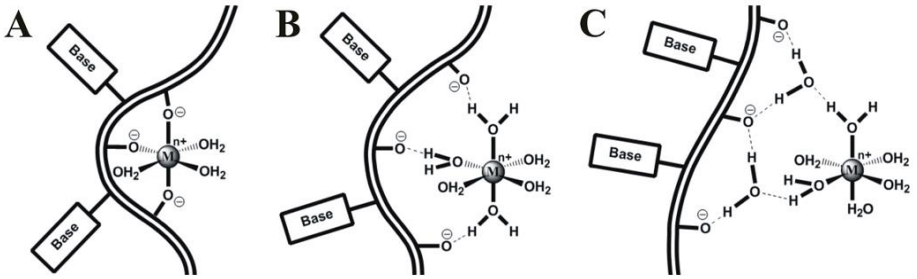



**Figure 2-4**. Modes of metal ion binding to RNA. Cations ($Mn^+$) can interact with RNA (A, B) requiring at partial and total dehydration of the metal ion, or transient, long-distance interactions between the solvated RNA and metal ion (C) (Johnson-Buck et al. 2011).

Divalent cations, mostly $Mg^{2+}$, $Ca^{2+}$, $Zn^{2+}$, $Fe^{2+}$ and $Cu^{2+}$, together with RNA/DNA may have played an important role in the catalytic reactions relevant to the synthesis of prebiotic macromolecules during the early evolution of life. Even in the modern biology world, one in third of enzymes known so far are metalloenzymes (Holm et al. 1996). Magnesium, the second enriched divalent cation in the biosphere, almost dominates the biological system. Among the common divalent cations, $Mg^{2+}$ has small ionic radius ($Mg^{2+}$ = 0.72Å, $Ca^{2+}$ = 0.99Å, $Na^{+}$ = 0.95Å, $K^{+}$ = 1.52Å), a high charge density and high solubility at neutral pH (Bowman et al. 2012). It can efficiently neutralize the negative charge of the electron donors, like oxygen, nitrogen and sulfur atoms (**Figure 2-4**). The involvement of magnesium in almost all biochemical processes related to translation is now well established (Lightfoot 1988; Lancaster et al. 2006; Sun and Zhang 2008): From the tRNA folding and tRNA aminoacylation to ribosomal structure and activity, magnesium directly or indirectly interact with substrates to promote catalysis (al-Karadaghi et al. 1996; Banik and Nandi 2010; Bhaskaran et al. 2012; Petrov et al. 2012). In relation with the ATP activation issue, the catalytic site of Lysyl-tRNA class II synthetase (**Figure 2-5**) requires three $Mg^{2+}$ to present in the amino acid adenylation step by bending the triphosphate groups of ATP into U-shape to expose the $\alpha$-phosphate for nucleophilic attack (Desogus et al. 2000). On the other hand, a class I aminoacyl-tRNA synthetase such as Tryptophanyl-tRNA synthetase requires only one $Mg^{2+}$ for amino acid activation (Retailleau et al. 2007). The presence of $Mg^{2+}$ ion located near the $\beta$ and $\gamma$ phosphate linkage can significantly reduce the negative charge density of the oxygen atoms and thereby aid the nucleophilic attack by the carboxylic group of the amino acid. Resent molecular dynamics simulations research also shows that $Mg^{2+}$ can support ribozyme structural stability and possibly facilitating catalysis (Ucisik et al. 2016). In our selection experiment, 5 mM $MgCl_2$ are included in the buffer to facilitate potential aminoacylation.

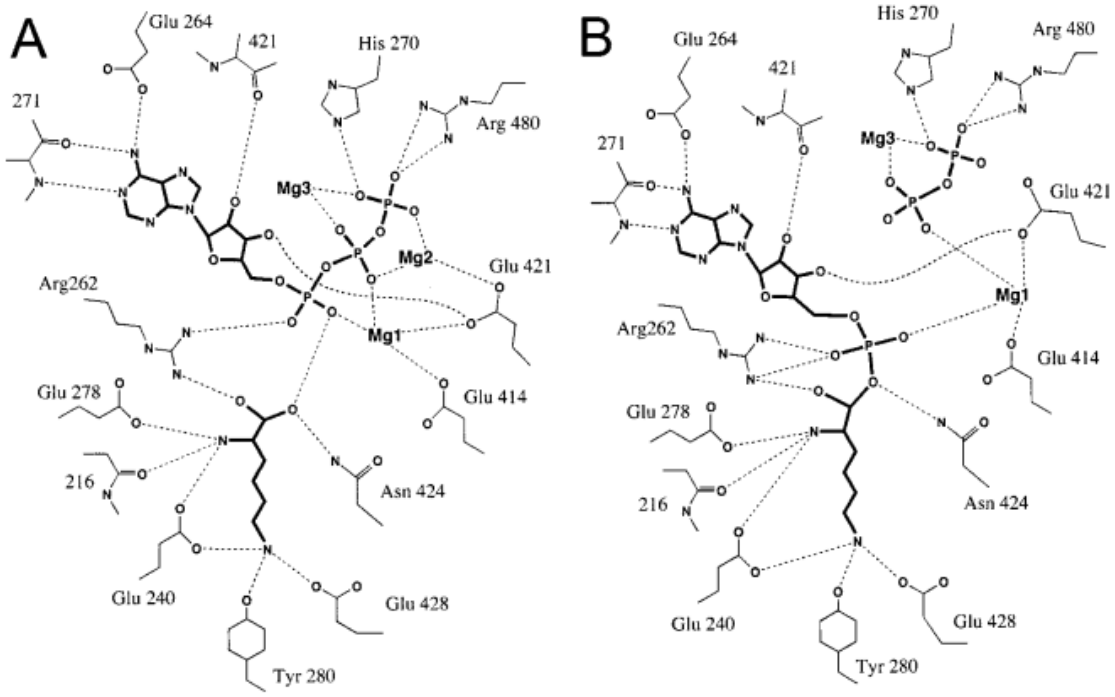



**Figure 2-5**. Schematic representation of the active site of LysU aaRS (*E. coli*), showing predicted hydrogen bonds (A) in the ternary complex with lysine and ATP and (B) in the complex with the lysyl-adenylate intermediate. (Desogus et al. 2000).

Calcium is the most abundant divalent cation and has a moderate charge density compare to $Na^+$ and $Mg^{2+}$. Because of its large radius and coordination distance with ligands, calcium usually isn't compatible with phosphoryl transfer reactions (Yang et al. 2006). Exceptionally, the presence of calcium alone is indeed able to catalyze amino acid activation with the RNA 5'-end triphosphate under an acidic pH condition by ribozyme KK13 (Kumar and Yarus 2001). In another case, it also has been reported that $Ca^{2+}$ may efficiently inhibit the first splicing step of group II intron (Erat and Sigel 2008). Perrota and coworkers have observed an interesting reactivity switch between genomic and antigenomic HDV ribozyme by changing the preference for $Mg^{2+}$ and $Ca^{2+}$ respectively, which means the genomic HDV ribozyme structure cleaves a bit faster in $Mg^{2+}$ than in $Ca^{2+}$ while the antigenomic form acts in a contrary way (Perrotta and Been 2007). These various properties of calcium indicate that it must have been played a critical role during the evolution. A reasonable concentration of $Ca^{2+}$ was included in the cocktail buffer to stabilize the ribozyme structure and possibly play a role in the activation of the amino acids.

Another well-studied divalent cation is manganese ($Mn^{2+}$). This ion is the most similar to $Mg^{2+}$ ($Mn^{2+}$ radius is 0.70Å). Experiments of $Mn^{2+}/Mg^{2+}$ replacement are often conducted to

understand the catalytic mechanism of ribozymes. Sometimes the effects could be dramatic (Young et al. 1997; Pontius et al. 1997; Schnabl and Sigel 2010). The smallest and simplest self-cleavage ribozyme was discovered as a manganese-dependent RNA whose activity is not affected by the presence of magnesium (Dange et al. 1990). The very similar small ribozyme present in the 3'-UTR of Vg1 mRNA and *β*-Actin mRNA is also manganese-dependent, and $Mn^{2+}$ can only replaceable by cadmium (Kolev et al. 2008). In the case of the hammerhead ribozyme, the substitution of $Mg^{2+}$ to $Mn^{2+}$ dramatically promotes catalysis up to 400 fold (Roychowdhury-Saha and Burke 2006) and also exceptionally facilitate folding compared with other alkaline earth ions (Boots et al. 2008). It may participate in stabilizing the ribozyme folding and forms tighter ions binding site that can enhance the loop-loop interactions (Kisseleva et al. 2005). At the experimental level, $Mn^{2+}$ is usually at a concentration of millimolar range. Considering the presence of $Mg^{2+}$ and $Ca^{2+}$, we choose to limit the concentration of $Mn^{2+}$ to a micro-molar concentration.

### 2.3.2 Substrates for ribozyme aminoacylation: amino acids and activating cofactors

Considering the two chemical steps leading to the aminoacylation of an RNA (eqs. 1 and 2), the activation of the carboxyl group of the amino acids (eq. 1) represents a major barrier to overcome. In the modern genetic system, this activation is achieved with ATP

$$ATP + aa \leftrightarrow aa\text{-}AMP + PPi \quad \text{(1)}$$

$$RNA + aa\text{-}AMP \leftrightarrow aa\text{-}RNA + AMP \quad \text{(2)}$$

Due to the hydroxyl group being a rather poor active group, almost all successful SELEX investigations on self-aminoacylating ribozymes so far used pre-activated amino acid to bypass the difficult step of activation. Commonly employed leaving groups are AMP, CoA (Coenzyme A), CME (Cyanomethyl ester), CBT (*p*-chlorobenzyl thioester), DBE (*3,5*-dinitrobenzyl ester). Using these leaving groups, the activated forms of the amino acids are all produced by organic synthesis (Zhang and Cech 1997; Murakami et al. 2002; Li and Huang 2005; Niwa et al. 2009; Morimoto et al. 2011).

The *in vivo* activation of carboxyl group is usually achieved by either of three intermediates: thioesters, acetyl CoA and acyl phosphates. All the aminoacyl tRNA synthetases (aaRS) require ATP as activating cofactor for *in vivo* aminoacylation of the tRNA. Investigations of aaRS crystal structures have provided many insights into the mechanisms of enzymatic amino acid activation (Giegé and Springer 2012). Usually one to three $Mg^{2+}$ ions cooperate with amino acids side-chains in the catalytic site bend the triphosphate of ATP to expose the $\alpha$-phosphate to the acyl group of the amino acid substrate. Concurrently, ions and other positively charged functional groups act as strong Lewis acid to reduce the electron density of the triphosphate in order to facilitate the nucleophilic attack by the acyl group (**Figure 2-6**).

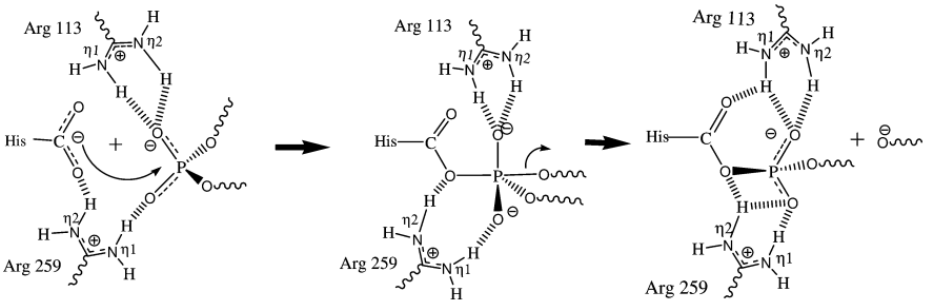



**Figure 2-6**. Schematic showing three stages of hydrogen bonding interaction of the active site residues with the reacting substrates of His-aaRS from *E. coli*. Reactants (left), intermediate (middle), and activated amino acid (right) (Banik and Nandi 2010).

Because RNA does not have positively charged residues, it is not clear how a ribozyme could achieve this activation; the only known activating ribozyme (Kumar and Yarus 2001) has not been structurally characterized so far. Among possible strategies are the coordination through divalent ions and protonated cytidine residue(s).

Based on few reported ribozyme crystal structures or computative structures, ribozyme can fold into complex tertiary structure motif with loops, stems, bulges, junctions, twisted bases to support multiple binding sites of ligands, such as ATP, AMP, amino acid, Adenine, NAD, CoA, etc. (Chen et al. 2007). ATP could have played an important role in ribozyme catalysis metabolism. If a ribozyme could catalyze aminoacylation, it firstly must provide appropriate binding pockets, such as amino acids, ATP and/or ions. *In vitro* selection has already illustrated such possibilities. Known ATP aptamers can either bind the adenine group (Dieckmann et al. 1996) or strongly recognize the triphosphate group (Sazani et al. 2004). Their binding

mechanism are all based on hydrogen bonds and RNA's folded structure (Tang and Breaker 1998). Another natural selected cofactor GTP is an essential substrate for group I intron *in vivo*. *In vitro* experiments have shown that if GTP is replaced by ATP, the first step splicing reaction can be inhibited, which indicates the GTP binding specificity (Raghavan et al. 2009). Compare with ATP, the guanidine group can provide one more hydrogen bond binding site and increase the affinity to RNA molecule. In our experiment we chose both ATP and GTP as activation cofactors with 2 mM as the final concentration.

Yarus and co-workers have analyzed the properties of various RNA-bound amino acids from riboswitches, aptamers and RNPs (Yarus et al. 2009). Different charges and sizes of side chain could guide to different binding mechanism and efficiencies. A few base mutations can change the structure of binding motifs and make them recognize different amino acids (Famulok 1994; Geiger et al. 1996). Moreover, amino acids themselves as cofactors could potentially participate in the catalysis (Roth and Breaker 1998). In order to offer a maximum of possibilities to the selection, we choose to work with a mixture of 19 different amino acids during the incubation step.

### 2.3.3 Other cofactors

Short polymers accumulation may have helped macromolecules to gradually construct the prebiotic world (Pressman et al. 2015). Besides RNA, other polymer chains, such as TNA, PNA, GNA and PDB, were also thought to be part of the original environment (Joyce 2002; Yakhnin 2013). All these hypothetical polymers could also have undergo a co-evolution process (Higgs and Lehman 2014).

Some of these polymers are known to affect the activity of RNA molecules (Stolze et al. 2001). In the cellular environment, RNA is mostly present in the cytoplasm, where up to 30% in volume is occupied by different macromolecules and crowders. *In vitro* experiments have shown that uncharged polymers such as polyethylene glycol (PEG), can either stabilize folded RNA structures or in another case decrease the stability of hammerhead ribozyme stem helices (Nakano et al. 2009; Kilburn et al. 2013). Depending on PEG molecular weight and ion

concentration, the hammerhead ribozyme can exhibit different activities due to a change in cation binding properties (Karimata et al. 2006; Nakano et al. 2008). On the other hand, charged polymers like peptides with functional side-chains could also facilitate the activity of RNA in catalysis. These polymers could possibly protect RNA from degradation, promote loops closures and even strengthen double-helices (**Figure 2-7** A) (van der Gulik and Speijer 2015; Carter 2015). Recently, using group II intron as a model to mimic the *in vivo* condition, Firorini and co-workers have found that crowding environment can affect RNA folding and activity (Fiorini et al. 2015). Molecular crowding was also found to help a mutated ribozyme to overcome destabilization (Lee et al. 2015).

Peptides have already been used in ribozyme selection to promote the aminoacylation. The RNA world hypothesis does not imply that RNA could perform any primordial task without cofactors such as small peptides, and both ribozymes and cofactors may have co-evolved in the prebiotic world (Bashan and Yonath 2005; Bowman et al. 2015).

As additional ingredients in our aminoacylation buffer, Poly-L-lysine, Poly-L-arginine and spermine (**Figure 2-7** B) are added as cofactors at low concentration to provide positive charges and potential cooperativity that may help RNA catalysis.

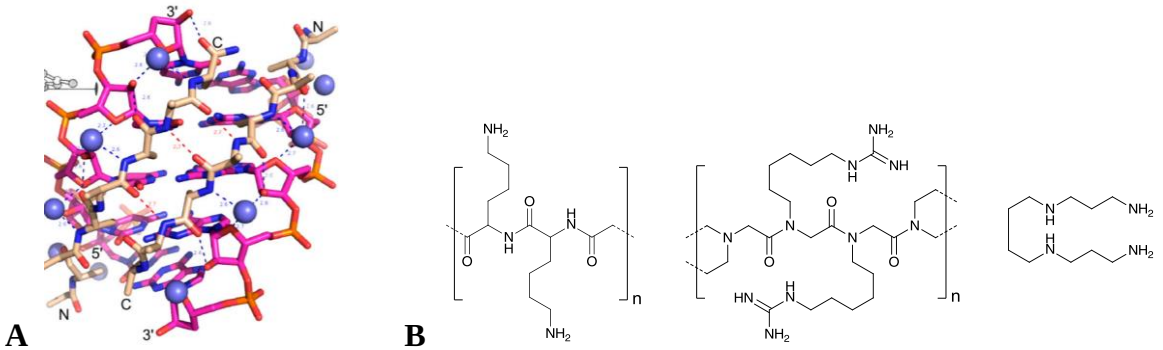



**Figure 2-7**. **A**: Stereochemistry of peptide-RNA construction. The peptide (inside) binds the RNA molecule (outside) into double helixes and stabilized by the hydrogen bonds between peptide carbonyl and the ribose 2'-OH groups, between amide nitrogen and water molecules (blue spheres) between the ribose $O_1$ and 2'-OH groups (Carter 2015). **B**: poly-L-lysine, poly-L-arginine and spermine.

### 2.3.4 Physico-chemical conditions: pH, incubation time and temperature

The choice of an appropriate pH is certainly critical to the success of our experiments. Although most enzymatic reactions occur in neutral or slightly alkaline pH, it is no clear what were the prebiotic pH conditions. Since molecular self-assembling started billion years ago under an atmosphere enriched in $CO_2$, ocean were likely acidic (pH 3.5-6), and natural selection could possibly begin in this environment, where RNA, NTPs, peptide as well as some biological micro molecule are rather stable (Bernhardt and Tate 2012).

Ribozyme *in vitro* selection has been mostly operated under neutral or slightly acidic pH. However, the only reported ribozyme KK13 that is able to catalyze amino acid activation best functions at pH 4 (Kumar and Yarus 2001), while the catalysis of transesterification by ribozymes is optimal at a pH around 7.0 (Chumachenko et al. 2009). On one hand, tRNA and most of the ribozyme catalyzed aminoacylation requires slightly alkaline pH for their functions; on the other hand, acidic pH can prolong activated amino acid half-life and prevent aminoacyl ester bond hydrolysis (Ninomiya et al. 2004). In order to best cope with these contradictory requirements, we implemented switch-pH conditions in our experimental design. The final concentration of the amino acids mixture was set to 17 mM in our experiments. Without any pH adjustment, the pH of the incubation buffer would be around 1-2. By using NaOH to carefully increase the pH to 3.5 and incubate overnight on ice, any potential amino acid activation would be stabilized under these conditions. Then, the pH was increased to 7.0 by the addition of an appropriate volume of HEPES buffer, and the solution was further incubated on ice for 2 hours to accumulate any possible aminoacylation events. The fully established protocol is shown in **Table 2-6**.

**Table 2-6**. Protocol of incubation to promote potential ribozyme aminoacylation

| Components | Volume | Final concentration |
|---|---|---|
| RNA (5 $\mu$M) | (adjustable) 5 $\mu$l | 0.25 $\mu$M |
| $H_2O$ | (adjustable) 5 $\mu$l | |
| NaCl (1 M) | 10 $\mu$l | 100 mM |
| KCl (1 M) | 10 $\mu$l | 100 mM |
| Total | 30 $\mu$l | |
| Incubate at 90°C for 3 min and cool down to 60°C | | |
| $CaCl_2$ (250 mM) | 1 $\mu$l | 2.5 mM |
| $MgCl_2$ (250 mM) | 2 $\mu$l | 5 mM |
| $MnCl_2$ (250 $\mu$M) | 2 $\mu$l | 5 $\mu$M |
| Poly-lysine (0.01%) | 2 $\mu$l | |
| Poly-arginine (0.01%) | 2 $\mu$l | |
| Spermine (250 $\mu$M) | 1 $\mu$l | 2.5 $\mu$M |
| aa Mix (42.5 mM) | 40 $\mu$l | 17 mM |
| Mix well and chill on ice | | |
| GTP (100 mM) | 2 $\mu$l | 2 mM |
| ATP (100 mM) | 2 $\mu$l | 2 mM |
| Mix well and avoid any precipitation | | |
| NaOH (1 M) | 3.5 $\mu$l | 35 mM<br>pH = 3.5 |
| Total | 88 $\mu$l | |
| Incubate on ice overnight | | |
| HEPES buffer (1 M, pH=7.0) | 12.5 $\mu$l | pH 7.0 |
| Total | 100.5 $\mu$l | |
| Incubate on ice for 2h and then precipitate | | |

## 2.4 Selection procedure and controls: the OD-DO Assay

### 2.4.1 Sodium periodate oxidation

Sodium periodate ($NaIO_4$) oxidation has been widely used in organic chemistry for specifically oxidizing *cis*-diol groups into di-aldehyde group. *Cis*-diol groups are normally present on the 3'-end of most RNA molecules. Under appropriate reaction conditions, periodate can eliminate the ribose five-membered ring up to 99% (**Figure 2-9**) (Hughes and Nevell 1948; Loring and Levy 1956). When this oxidation is followed by an incubation at alkaline pH ($\geq$ 9.5), it leads to a so-called $\beta$-elimination: the terminal oxidized nucleotide of an RNA gets removed from the

chain, and yields a terminal phosphorothioate group (**Figure 2-9**) (Alefelder et al. 1998; Akbergenov et al. 2006; Kawano et al. 2012).

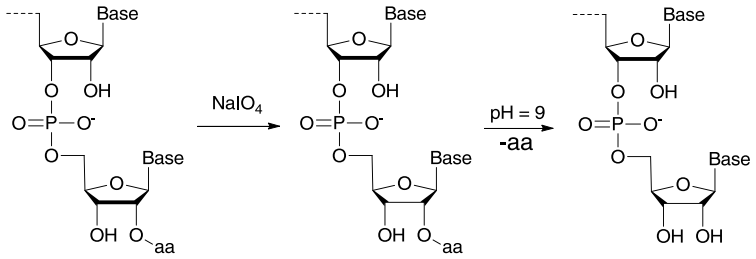



**Figure 2-8**. Schematic diagram shows showing the result of an **OD** procedure.

Many previous studies have introduced sodium or potassium periodate to oxidize single nucleotides (Sufrin et al. 1995) or the 3'-end of RNA molecules (Dittmar et al. 2005). Here the periodate is commonly used as a discriminative reagent to distinguish the differential of modified or non-modified RNA 3'-end (Kurata 2003; Dittmar et al. 2005; Behm-Ansmant et al. 2011). When followed by *β*-elimination (**Figure 2-9**), periodate oxidation can be used to generate a terminal phosphorothioate or monophosphate group on RNA that prevents intramolecular cyclization or multiple additions during RNA ligation (Sninsky et al. 1976; Uhlenbeck and Cameron 1977; Bruce and Uhlenbeck 1978; Schutz et al. 2010).

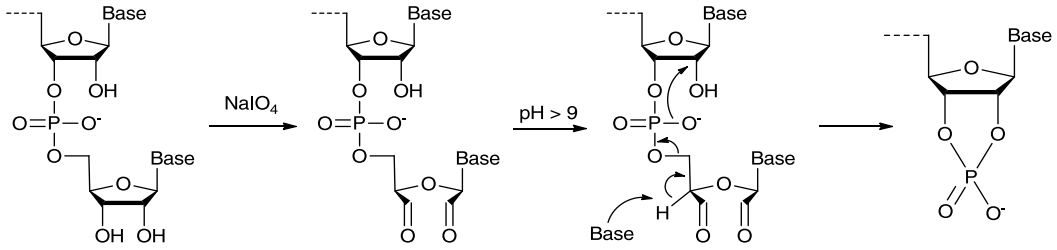



**Figure 2-9** The mechanism of sodium periodate oxidation and *β*-elimination.

### 2.4.2 The OD-DO Assay

The selection of aminoacylated RNA was established with so-called Oxidation-Deacylation and Deacylation-Oxidation (**OD-DO**) assays. These assays allow us to select aminoacylated RNA (**OD**) and also give a negative control (**DO**) to the procedure. Two additional positive controls (**D** and **NT**, no treatment) were also implemented. A similar protocol known as OXOPAP assay

has been formerly used by Dittmar *et al.* and Gaston *et al.* to quantify aminoacylated tRNA (Gaston et al. 2008; Puerto-Galán and Vioque 2012; Dittmar et al. 2005). We use the same strategy with optimized condition to select our targeted ribozymes. RNA pools that underwent an aminoacylation assay (**Section 2.3.5**) are split into different batches for different oxidation and/or deacylation treatment (**Figure 2-10**)

1) In one batch of experiments, we first treat the RNA sample with $NaIO_4$. Oxidation followed by deacylation (**OD**) will allow us to select aminoacylated RNA, and thus any potential ribozyme.

2) On the contrary, deacylation followed by oxidation (**DO**) is a negative control to evaluate the efficiency of a complete elimination of the RNA from the selection procedure.

3) A positive control: Deacylation only (**D**).

4) An additional positive control: No Treatment (**NT**).

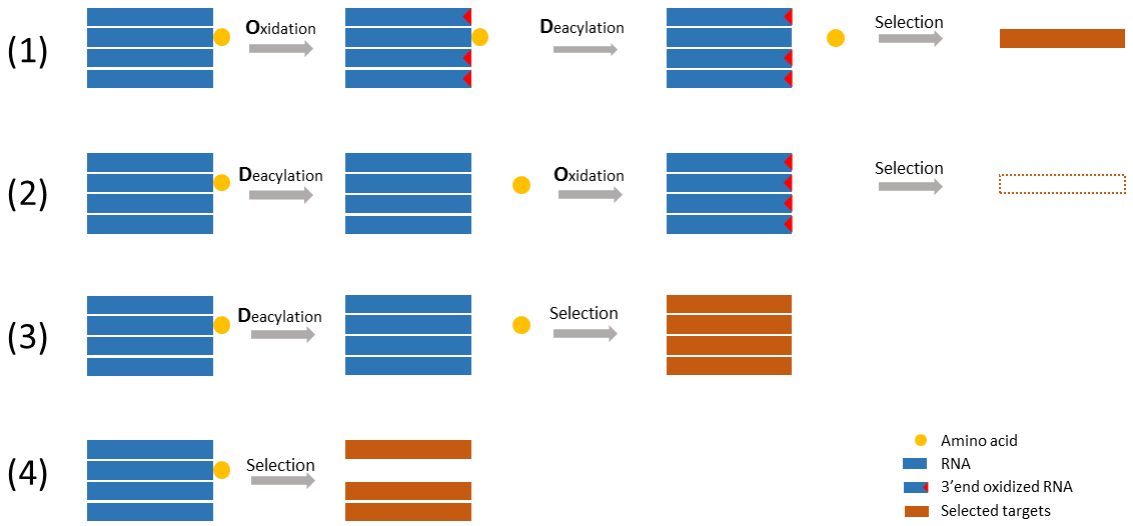



**Figure 2-10**. Outline of OD-DO assay. (1), (2), (3), (4) stand the experiment workflow of **OD**, **DO, D, NT**, respectively.

While repeating the OXOPAP assay (Gaston et al. 2008), we found that the saturating concentration of $NaIO_4$ used by them (0.5 M) is not an appropriate in our protocol. It turns out that the leftover sodium periodate cannot be removed by precipitation: because neither sodium periodate nor its reductive product, sodium iodate ($NaIO_3$), is soluble in ethanol, the usual ethanol precipitation procedure causes a mass of flocculent co-precipitation which tightly binds to the RNA molecules. Although a much lower concentration of $NaIO_4$ was finally used in our

oxidation experiments, we realized at a late stage that it could still prevent the downstream enzymatic reactions and generate false negative results. In our **OD-DO** assays, the **DO** control would usually show the expected (negative) signal. However, in this control, oxidation precedes the ligation of a 3' RNA adapter, which is the next critical step in our protocol (**Section 2.5**). It turned out that the leftover periodate could still significantly hamper RNA ligation, and thus prevent a final RT-PCR amplification. This phenomenon was discovered during our investigations on total RNA (**Chapter II**): **Figure 2-11** shows the analysis of total RNA from *E. coli* on a 3-4% agarose gel. After an **OD-DO** assay treatment, approximately 800 ng of total RNA from each experiment group was loaded on gel (**Left**). Almost no RNA is visible in the **DO** lane. However, after a 7 h treatment at 37°C with DNase I followed by phenol/chloroform extraction, the same RNA sample can be perfectly seen on the gel (**Right**). It indicates that non-removed sodium periodate or sodium iodate can trap the RNA molecules due to the co-precipitation and possibly affect the downstream enzymatic reactions. This result however shows that a long incubation in appropriate conditions can re-nature the RNA.

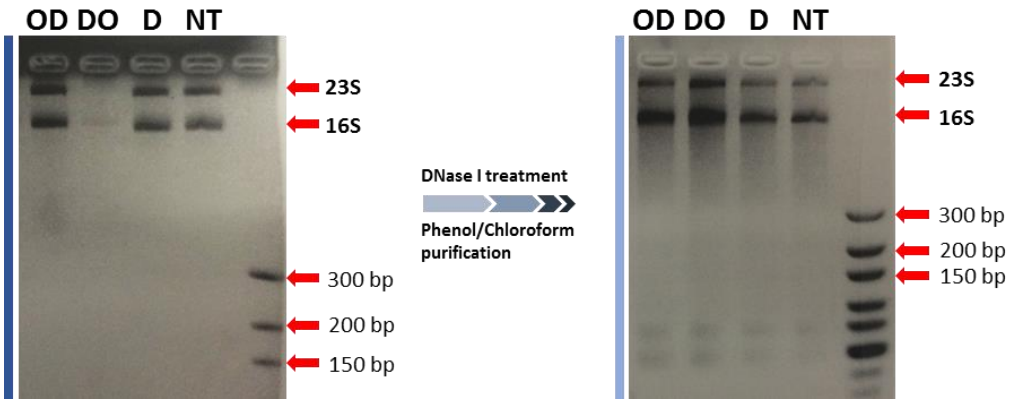



**Figure 2-11**. Gel analysis of total RNA. Left gel shows the total RNA loaded on agarose gel after OD-DO treatment. The lane "DO" shows an obvious disappearing of RNA. Right gel shows the same total RNA loaded on agarose gel after DNase treatment. The lane "DO" is showing up again.

All our oxidation experiments are performed in the dark to prevent the photochemical decomposition of periodate and periodate oxidation of formic acid by ultraviolet light (Marinetti and Rouser 1955). After oxidation, the excess of unreacted periodate is usually quenched with two equivalents of glucose to redox the leftover $NaIO_4$, and then a gel spin column is applied to remove most of the remaining $NaIO_3$ (Dittmar et al. 2005). Because the small size of our RNA molecules is incompatible with these columns, this column purification step is not

implemented in our experiment. Instead, we supply 10 mM of DTT to redox $NaIO_3$ to NaI (Hughes and Nevell 1948; Ivery et al. 1984). A final ethanol precipitation efficiently eliminates the oxidant.

The deacylation conditions also need to be carefully established. Most RNA molecules are unstable in alkaline pH solutions, especially when cations such as $Mg^{2+}$, $Zn^{2+}$, $Ca^{2+}$ are present. In a pH higher than 9, RNA will start to degrade (Kawano et al. 2012). On the other hand, if the pH is not high enough, deacylation is inefficient, especially in a buffered solution without ions. We have investigated several conditions and time-course experiments while varying pH, temperature, incubation time, and PCR cycles, (data not show) to optimize the deacylation condition.

### 2.4.3 Positive control

In order to validate the oxidation and deacylation procedures, our initial tests were realized with several chemically synthesized tRNA fragments of about 50 nt (we call them mini-tRNA) from *Staphylococcus aureus* (HG003) $tRNA^{Ala}$, $tRNA^{Val}$, and $tRNA^{Gly}$. The 3'-end of these RNA is not aminoacylated, and should thus be oxidized. All our initial oxidation tests failed to prevent the appearance of a final PCR product (30 PCR cycles are usually performed). $NaIO_4$ cannot efficiently oxidize these RNA, and the adapter ligation still works sufficiently enough to enable RT-PCR. **Figure 2-12** shows an oxidation experiment results.

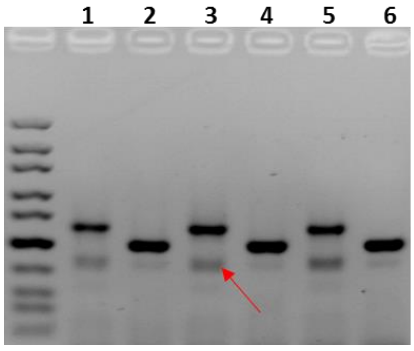



**Figure 2-12**. The final RT-PCR result of mini-tRNA Ala (1, 2), Val (3, 4) and Gly (5, 6) after treat with $NaIO_4$. Clearly, after oxidation the lane 1, 3, 5 should have shown a negative result of RNA ligation. Lane 1, 3, 5 represent the PCR products which generate from the ligation of mini-tRNA and RNA adapter. Lane 2, 4, 6 represent the PCR product of mini-tRNA themselves (Ala, Val and Gly) as a control. Red arrow points the primer-dimer by-product.

We suspected that the chemically synthesized RNA may still keep some protecting group at the 2' or 3'-end of the RNA that could prevent oxidation (Caruthers 2013). We therefore tried our oxidation experiments with *in vitro* transcribed mini-tRNA$^{Val}$. **Figure 2-13** shows that these RNAs respond much better to oxidation.

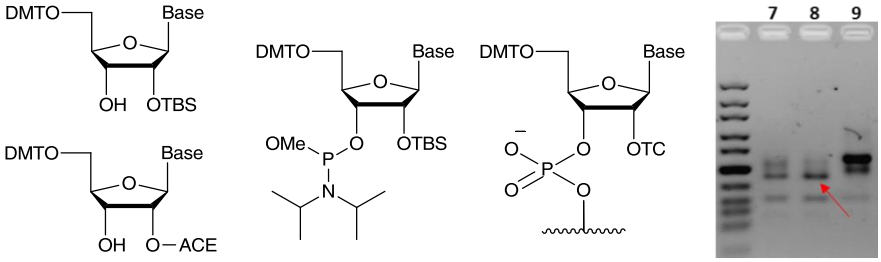



**Figure 2-13.** Left: the commonly used protecting group during chemically RNA synthesis. Right: lane 7 shows the $NaIO_4$ treated RNA has much lower ligation efficiency than the non-treated one (lane 9). Lane 8 shows a negative control which didn't do the RNA ligation. Red arrow points the primer-dimer by-products.

To test the sensitivity of the **OD-DO** combined assays, we used tRNA$^{Tyr}$ (85 nt) from *E. coli* as a positive control. Aminoacylation of tRNA$^{Tyr}$ is firstly performed *in vitro* by the specific aminoacyl-tRNA synthetase (Soutourina et al. 1999). Subsequently, the solution with aminoacylated tRNA$^{Tyr}$ is divided into 4 different aliquots. Oxidation and deacylation experiments are performed with tRNA$^{Tyr}$ at a final concentration of 0.08 $\mu$M. After **OD**, **DO**, **D** and **NT** treatments, a final RT-PCR is achieved with a primer complementary to the 3'-adapter and a primer complementary to the 5'-end of tRNA$^{Tyr}$. Results are presented in **Figure 2-14**. With 20 cycles of PCR, a significant difference is observed between **OD** and **DO** experiments. To confirm the sensitivity of this protocol, it was further applied on total RNA extracted from *E. coli* BL21 by acid phenol method (**Supporting information**). Acidic pH prevents deacylation during extraction. An amount of 4 $\mu$g total RNA was loaded into 4 tubes and went through identical **DO/OD/D/NT** treatments. To prevent genome amplification, all the samples were subsequently treated with DNase I at alkaline pH. The final 20 cycles PCR shows that a similar difference if observed between **OD** and **DO** experiments.

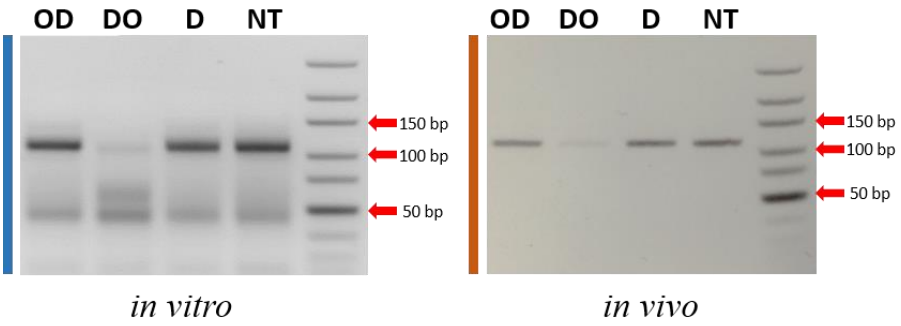



**Figure 2-14**. $tRNA^{Tyr}$ from *E. coli* is employed as positive control to verify the protocol. $tRNA^{Tyr}$ is a specific 85 nt long natural tRNA. After **OD-DO** assay treatment, it will subsequently ligate with a carefully designed RNA adapter P with 23 nt in length. By using $tRNA^{Tyr}$ 5'-reverse primer and primer of the adapter, the expect products should show a size at 108 bp. Left: Aminoacyl $tRNA^{Tyr}$ yield by *in vitro* aminoacylation with commercial purified natural *E. coli* $tRNA^{Tyr}$, Tyrosine, and *E. coli* Tyrosine-specific aaRS. Gel shows the 20 cycles PCR product of *in vitro* selected aminoacyl $tRNA^{Tyr}$. 4 lanes represent **OD/DO/D/NT** treatment. Right: 1 $\mu$g total RNA extracted from *E. coli* at $OD_{600}$ 0.4 is treated by **OD/DO/D/NT** procedure. Gel shows the final 20 cycles PCR result.

These two positive controls show that our **OD** protocol can efficiently remove non-acylated RNA from either specific pools of RNA or from total RNA. The detailed **OD-DO** protocols are shown in **Table 2-7**.

**Table 2-7**. OD-DO assay

| | RNA pellet | | | |
|---|---|---|---|---|
| | OD [a] | DO [a] | D | NT |
| NaOAc/HOAc (3M pH=5.2) | 5 $\mu$l | 0 | 0 | |
| $H_2O$ | 40 $\mu$l | 43 $\mu$l | 43 $\mu$l | 50 $\mu$l |
| $NaIO_4$ sol. (0.1M) [b] | 5 $\mu$l | 0 | 0 | |
| Borax buffer (pH=10.0) | 0 | 7 $\mu$l | 7 $\mu$l | |
| | on ice 45 min [c] | 42ºC 1h | 42ºC 1h | on ice 1h |
| | | Precipitation [c] | | |
| NaOAc/HOAc | 0 | 5 $\mu$l | | |
| $H_2O$ | 43 $\mu$l | 40 $\mu$l | | |
| $NaIO_4$ sol. (0.5M) [b] | 0 | 5 $\mu$l | | |
| Borax buffer (pH=10.0) | 7 $\mu$l | 0 | | |
| | 42ºC 1h | on ice 45 min [c] | | |
| | | Precipitation [c] | | |

[a] All the tubes covering with aluminum films for oxidation

[b] Dissolve 5.4 mg $NaIO_4$ in 250 $\mu$l water to prepare 0.1 M sodium-periodate solution

[c] After oxidation, add 10 $\mu$l Glucose (0.5 M) and incubate 15 min on ice. Then add 1 $\mu$l Glycogen as well as 6 $\mu$l DTT (0.1 M) into the oxidation solution.

## 2.5 RNA ligation

### 2.5.1 Adapter design

After the **OD** treatment, only selected RNA that keep normal *cis*-diol group at the 3'-end could be ligated to a 5' phosphorylated RNA adapter with a T4 RNA ligase 1 (Rnl1). Nine different RNA adapters were successively tested to optimize the ligation reaction and the subsequent RT-PCR protocols. They can be categorized into 4 classes based on their sequence and/or structure profiles (**Figure 2-15**). In the first class, all adapters are single stranded. This class includes adapter J, K, L, XL, and M (former tests with adapter A to I were performed by Maser student Lauriane Cacheux). All other adapters are RNA-DNA hybrids. Among those RNA-DNA hybrid adapters, class II and class IV are similar except for the length of the single stranded portion of the RNA strand and the restriction site, while class III adapter includes a partially non-complemented DNA strand. These chimeric adapter molecules (class II to IV) were designed to prevent unspecific priming on the random track during reverse transcription (See below).

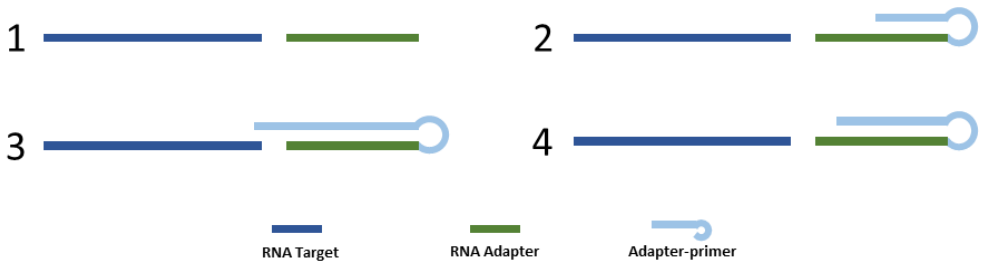



**Figure 2-15**. **1:** 5' phosphate adapters tested during the establishment of the protocol. Class I adapters are all single stranded RNA adapter, including adapter J, K, L, XL, and M; **2**: Class II adapter is an RNA/DNA hybrid, that includes adapter Hyb-XL; **3**: Class III adapters have similar structure to class 2 except the hybrid DNA part has an overlap with the constant 3'-end of the targeted RNA, such as adapter N and NR; **4**: Class IV adapters are similar to class 2 but carry a different restriction site. It includes adapter O and P. Sequence details see **Table 2-8**.

RNA ligation plays an important role in this protocol. Adapters are working as fishing reagent to "harvest" the interesting target out of the mixture. RNA ligation, especially by Rnl1, is an enzymatic reaction with low efficiency, with observed ligation rates spanning from 0 to 80% (Turunen et al. 2014). Ligation experiments with very short (1 to 5 nt) RNA fragments have shown the sequence ApApA to be a much better acceptor substrate than UpUpU. Moreover,

ApA is less active than ApApA (Kafumann and Kallenbach 1975; Sugino et al. 1977). It has also been shown that short poly A acceptors and poly U donors form a pair with good ligation efficiency (Ohtsuka and Nishikawa 1976). Extensive ligation experiments with random 21-nt RNA molecules have revealed that less than three un-structured nucleotides at the 3'-end and RNA are likely to be poorly ligated (Zhuang et al. 2012). Furthermore, a 3'-end adenosine is preferred over cytidine and guanosine for ligation, while a uridine residue is a relatively poor acceptor substrate (A > C ≥ G > U) (England and Uhlenbeck 1978; Romaniuk et al. 1982). Our own tests with random pools designed with constant 3'-end showed that a much higher ligation efficiency is observed with "UpCpA" compared with "UpCpG" ending (**Section 2.2.1**). On the donor side, experiments with minimal donor substrates showed a decreasing ligation efficiency with the order pCp > pUp ≈ pAp > pGp (Romaniuk et al. 1982). All the tested adapters we designed have a short 5'-end poly A/U tail with a terminal phosphate group.

**Adapter J, K, L, XL and M**

Adapters K, L and M (**Table 2-8**) were the first to be investigated. They are all single stranded, with a short poly-U tail at the 5'-end. Adapter K was originally designed to test RNA ligation with an *in vitro* transcribed $tRNA^{Val}$ from *Staphylococcus auras*. Since tRNA are usually well folded molecules with a short unstructured CCA ending, ligation with a single strand RNA adapter works quite efficiently based on our experiment. Intermolecular ligation between adapters is however still observed in that case. Adapters L and M were designed to solve this problem; they contain an inverted dT oligo at the 3'-end to prevent intermolecular ligation. Although this setup was found to work well with specific tRNAs, it turns out that side-products are always present while working with pools of random RNA. Clone sequencing revealed that these short PCR side-products are primer dimers (primer of the adapter + T7 promoter primer) with a small stretch of bases in between, indicating that unspecific priming occurs on the random region, either during reverse transcription or during PCR (**Figure 2-16**).

In an attempt to avoid this unspecific product, we extended the length of RNA adapter (**Figure 2-16**, light green) and the DNA anchor (**Figure 2-16**, light orange) up to 32 nt (we called it Adapter XL). With an annealing temperature of 72°C, we hoped that while using a two steps hot-start PCR procedure, this would prevent the unspecific amplifications, but it did not work.


**Figure 2-16**. The three colors line represent cDNA (3'→5'): Light green is the primer of the adapter part initially come from the ligated RNA adapter; Light orange is the cT7 promoter sequence which is added by DNA ligation; Light blue is complementary DNA of the random RNA. Dark green is the primer of the adapter forward primer. Brown is T7 promoter reverse primer (5'→3').

**Table 2-8**. The sequences of Class I adapters (Grey color indicates the restriction site).

| Name | Sequence (5'→3') |
|---|---|
| Adapter K | 5Phos/rUrUrUrUrGrArArGrArGrCrGrGrCrCrG |
| Adapter L | 5Phos/rUrUrUrUrCrArGrGrGrArUrArCrGrCrCrGrCrU/3InvdT/ |
| Adapter XL | 5Phos/rUrUrUrUrCrArGrGrGrArUrArCrGrGrGrArGrUrUrArArCrU rUrUrGrCrArUrArGrGrCrGrArUrUrGrCrArA/3InvdT/ |
| Adapter M | 5Phos/rUrUrUrUrCrArGrGrGrArUrArCrGrCrArGrUrCrUrArCrU rG/3InvdT/ |

A control experiment showed that unspecific priming was a concern especially during reverse transcription: while performing an RT-PCR with RNA, RNA adapter and primer of the adapter without prior RNA ligation, the same primer-dimer byproducts were generated.

Because the RNA adapter is in excessed during ligation compared to the pool of random RNA, so is the primer of the RNA adapter during reverse transcription. Thus, unwanted associations between this primer and the random RNA region would always occur, either with or without mismatch(es) (**Figure 2-17**). Because these reverse transcription events are not limited by RNA ligation (as it occurs with our targeted selection), they become the dominant product of the final PCR. This issue was solved with the design of combined primer of the adapter molecules (see next Section).

**Figure 2-17**: Random priming during reverse transcription. Light blue lines represent primers of the adapter (3'→5'); The dark green line is the ligated RNA adapter (5'→3'); The dark blue line stands for the random RNA pool (5'→3'); The red arrow indicates the reverse transcription orientation.

**Adapter Hyb-XL**

Combined DNA/RNA hybrid adapters have been used as a molecular probes in some other sequencing protocol (Dittmar et al. 2005; Zheng et al. 2015). The main purpose of this design is that it may avoid any unspecific priming during the reverse transcription step. In order to work well, this strategy requires the hybrid double stranded adapter to be properly folded before ligation. It is achieved with a denaturation-renaturation procedure at low concentration.

**Table 2-9**. Sequence of Adapter Hyb-XL

| Name | Sequence (5'→3') |
|---|---|
| Adapter Hyb-XL | 5Phos/rUrUrUrUrCrArGrGrArUrArCrUrCrGrCrArGrUrUrArArCrUrCrGrGrCrArUrArGrGrCTTTTGCCTATGCCGAGTTAACTGC |

Adapter Hyb-XL is constituted by an RNA adapter linked to a DNA primer by a poly-dT bridge (**Table 2-9**). Once folded, this chimeric oligo has its primer base-paired to the RNA segment. Improper refolding could generate duplexes of these adapters (**Figure 2-18**), a structure that is not expected to affect RNA ligation and RT-PCR.

**Figure 2-18**. Correct folded adapter (Left) and adapter duplex (Right). The green part represents the RNA and the light blue part shows the complementary DNA primer as well as a poly-dT loop.

We investigated the annealing condition for this adapter. Unlike intermolecular annealing, intramolecular annealing requires diluted concentration and flash annealing**.** A rapid cooling procedure by dropping the solution into liquid nitrogen after a 90°C denaturing step was applied on different initial adapter concentrations from 0.1 to 10 $\mu$M. As expected, the 0.1 $\mu$M concentration shows the best sharp peak on bioanalyzer. **Figure 2-19**, part A, indicating the presence of essentially a single well-folded structure in solution. The signal is still good at 1 $\mu$M (**Figure 2-19**, B), but becomes bifurcated at higher concentrations, indicating the presence of another folded species. In order to be compatible with the RNA ligation protocol, we choose the 1.0 $\mu$M concentration to prepare our stock, which would avoid the need to re-concentrate the solution.

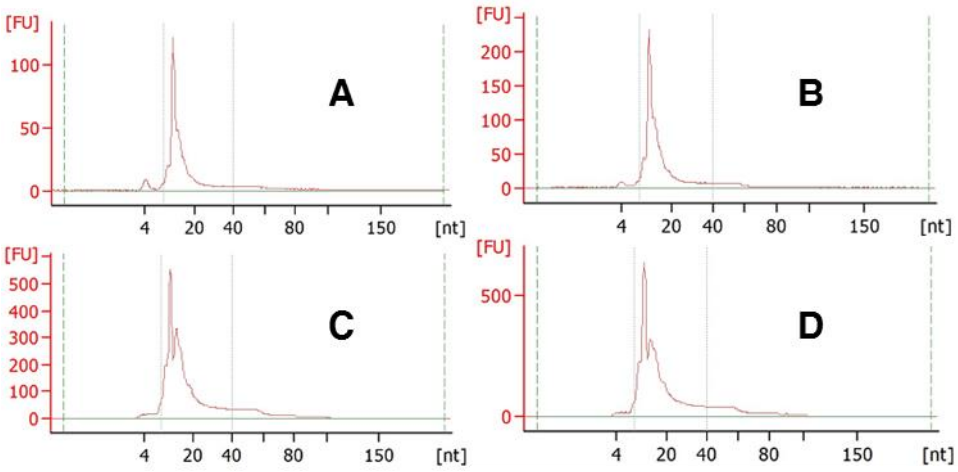



**Figure 2-19**. Bioanalyzer traces of annealed Adapter Hyb-XL in different concentrations ($\mu$M). A) 0.1; B) 1; C) 5; D) 10. Adapter Hyb-XL diluted in 100 $\mu$l water was incubated at a pre-heated PCR machine at 90°C for about 2 min and immediately dropped into liquid nitrogen. A sample volume of 0.1 $\mu$l was loaded on bioanalyzer (small RNA6000 chip).

The ligation protocol was tested with the pre-annealed Adapter Hyb-XL and the random RNA pool N37. The comparison of the bioanalyzer signal before and after ligation (**Figure 2-20**, resp. left and right) allows a clear identification of the ligation product (right, green arrow). A gel analysis of the RT-PCR amplification products was consistent with this result, and revealed both the ligated products and a by-product of a size similar to the two PCR primer-dimer. A $NaIO_4$ treated control sample mostly only showed a by-product close to the size of primer-dimer (data not show). These tests provided the first proof-of-principle of the design of our selection protocol.

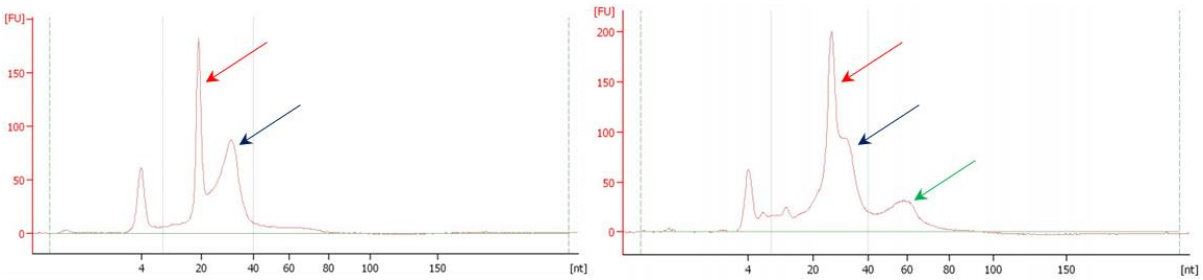



**Figure 2-20**. Bioanalyzer traces of RNA ligation between Adapter Hyb-XL and random RNA N37 before ligation (left) and after ligation (right). **Left**: 1 $\mu$M pre-annealed adapter mixing with random RNA N37 at a final molar concentration ratio 10:1. **Right**: Given such mixture, bioanalyzer shows the signal after RNA ligation at 4°C overnight. The red arrow represents the pre-annealed adapter Hyb-XL; The dark blue arrow shows the random RNA N37 pool; The green arrow indicates the expected ligation products.

An unwanted side effect however occurred with this Adapter Hyb-XL after reverse transcription. As a part of our initial design, we attempted to ligate a forward DNA anchor to the 3'-end of the cDNA (see **Section 2.7**), required to achieve a final PCR with essentially random pools (ref. **Figure 2-1**, step 6). The DNA anchor had a 5'-dGdG-3' sticky overhang to catch the terminal 3'-dCdC end of the cDNA. Coincidentally, the restriction site (**Table 2-9**, grey) of this adapter contains a 5'-rGrG-3' (red) sequence, corresponding to a 3'-dCdC-5' on the cDNA. Because some degradation of the single stranded part of the adapter could not be prevented, this 3'-dCdC-5' became available for ligation with the DNA anchor "5'-dGdG-3'" sticky end (**Figure 2-21**). Sequencing results allowed us to identify the primer dimer as well as this unexpected ligation product. This design was consequently abandoned.

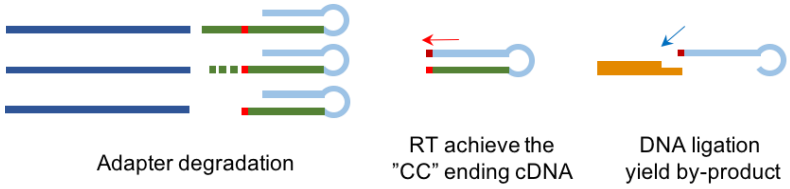



**Figure 2-21.** The main by-product ligation mechanism: the RNA tail of this hybrid adapter degrades to the "rGrG" part (red dot) and reverse transcribed cDNA carrying a 3'- "CC" ending (dark red) can ligate with the DNA T7 promoter anchor (orange). Dark blue represents RNA target, light blue is the DNA part of adapter, green shows the RNA part of adapter, red arrow indicates the reverse transcription orientation.

### Adapter N and NR

In another attempt to prevent to the appearance of PCR by-product, the class III adapters N and NR with a 3' overhang were investigated. Their sticky ends with 3 or 4 overhanging nucleotides were designed to bind the complementary 3'-end of targeted RNA molecules (**Table 2-10**, in red). A DNA ligase was used instead of T4 RNA ligase since the bound molecules form a duplex (Moore and Sharp 1992). A similar strategy was already applied to target tRNA with a TGG overhang (Dittmar et al. 2005).

**Table 2-10**. Sequences of Class III adapters

| Name | Sequence (5'→3') |
|---|---|
| Adapter N | /5Phos/**rUrUrUrU**rGrU**rGrGrArUrArC**rGrCrArGrUrCrUrArCrUr GTTTTCAGTAGACTGCGTATCCAC**AAAATGG** |
| Adapter NR | /5Phos/**rUrUrUrU**rGrU**rGrGrArUrArC**rGrCrArGrUrCrUrArCrUr GTTTTCAGTAGACTGCGTATCCAC**AAAATGGR** |

It turns out that another side-effect occurred with this design: the unstructured DNA tail of the adapter being free, unspecific priming was detected again that occurred during the reverse transcription step (**Figure 2-22**). An RT-PCR control performed while incubating the adapter together with a random RNA pool without DNA ligation would also yield the primer-dimer.

**Figure 2-22.** Adapter with 3' DNA overhang: unspecific priming during reverse transcription. Light blue: primer of the adapter; Green: RNA adapter; Dark blue: ligated RNA; Red arrow: reverse transcription orientation.

### Adapter O and P

Adapter O and P are the last two adapters we tested for RNA ligation, and are part of the final design of the selection protocol. These two adapters mostly overcome the problems encountered with the previous adapters: They can be efficiently ligated and enable specific RT-PCR. They differ in the restriction site (**Table 2-11**), and it turns out that only Adapter P could be efficiently cleaved after the final PCR of each selection cycle, which is required before a new RNA pool can be regenerated.

Adapter O was designed with a "BmrI" recognition site. This enzyme is a Type IIS restriction endonuclease that recognizes the asymmetric 6-bp sequence ACTGGG and cleaves 5 and 4 base pairs downstream on the up and down strand respectively [5'...ACTGGG$(N)_5$ /$(N)_4$…3'] (Chan et al. 2007; Bao et al. 2008).

**Table 2-11.** Sequences of Class IV adapters

| Name | Sequence (5'→3') |
|---|---|
| Adapter O | /5Phos/rUrUrUrUrUrCrCrCrArGrUrCrArGrCrUrGrUrCrUrArGrU<br>TTTTACTAGACAGCTGACTGGG |
| Adapter P | /5Phos/rArGrArArGrGrArGrCrCrGrUrUrArGrCrUrGrUrCrUrArGrU<br>TTTTACTAGACAGCTAACGGCTC |

This restriction site was selected to take advantage of the strong base paring interaction GGG/rCrCrC at the end of the duplex (**Figure 2-23**), which minimizes fraying (**Table 2-11**, red).

With this adapter, ligation and RT-PCR works much better than in all our previous attempts.


**Figure 2-23**. The sequence structure of Adapter O. Blue ball represents monophosphate group. Blue triangle stands by the restriction enzyme digestion site. In the red square circle, it is the restriction site.

Unfortunately, we could not achieve a proper digestion with BmrI already at the end of the first round selection. We tried to optimize temperature, substrate concentration, incubation time, and despites all our efforts it seemed impossible to get clear digestion products (**Figure 2-24**). Furthermore, the digested product was also very hard to recover by gel purification. The origin of this deficiency remained unclear, and we finally had to give up with this adapter and switch to Adapter P which contains a restriction site for the enzyme "EarI".

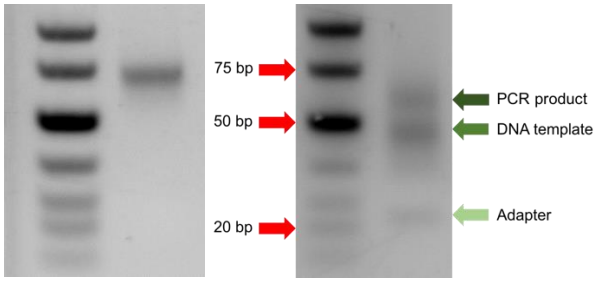



**Figure 2-24**. Gel analysis of before (left) and after (right) BmrI digestion of the PCR product after one round of selection. 0.5 $\mu$g PCR product (70 bp) was treated by BmrI FastDigest kit overnight. The green arrows indicate initial PCR product, digested DNA template and the removed adapter part, respectively.

EarI is a type II restriction endonuclease that enables to specifically cleave one nucleotide at 3'-end of the recognition sequence on one strand, and four nucleotides away from the 5'-end on the opposite strand [5'…CTCTTC(N)/($N_4$)…3'] (Polisson and Morgan 1988).


**Figure 2-25**. The sequence structure of Adapter P. Blue ball represents monophosphate group. Blue triangle stands by the restriction enzyme digestion site. In the red square circle, it is the restriction site.

Adapter P and Adapter O share similar sequence and melting temperature ($T_m$ = 75° C). A

high stability of Adapter P folding is made possible by rGrArGrCrCrG/CTCGGC base pairing near the 5'-end of the adapter (**Figure 2-25**). Instead of poly-U, an adenosine-rich 5'-end (rArGrArA), including some bases of the EarI restriction site is used for RNA ligation. It turns out that EarI could efficiently digest our PCR products, and provide new templates of sufficient quality. And finally, the EarI digestion can achieve to a good quality (**Figure 2-26**).

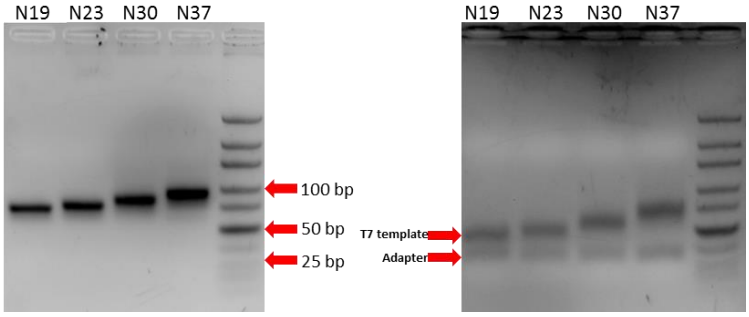



**Figure 2-26**. Gel analysis of before (left) and after (right) EarI digestion of the PCR product after one round of selection. 1 $\mu$g PCR product (4 libraries) was treated by EarI FastDigest kit (NEB) overnight. Red arrows indicate the digest products.

### 2.5.2 RNA ligation buffer

Macromolecular crowders, such as polyethylene glycol (PEG), polyvinyl alcohol (PVA) or polyvinyl pyrrolidine (PVP), have been shown to improve intermolecular ligation by increasing the relative concentration of both donor and acceptor ends through macromolecular crowding (Harrison and Zimmerman 1984; Munafó and Robb 2010). Besides, the ratio of accepter versus donner is also a critical parameter for RNA ligation. Quantitative ligation can be observed at low RNA accepter, with 15-20 times excess of donner molecules (Chumachenko et al. 2009). In our experiments, we also observed that a high concentration of RNA acceptor (up to 1 $\mu$M) strongly inhibits the ligation reaction.

In summary, RNA ligation with Rnl1 (ThermoFisher) should be performed at low concentration of RNA acceptor with an excessed of RNA donor. A crowed environment and low temperature will further improve ligation. Strong secondary structure as well as intermolecular aggregation will reduce the accessibility of the RNA termini, thus inhibiting the ligation reaction. In order to avoid this, the ratio of selected RNA and adapter is around 1/5 to 1/10 depending on the experiment.

## 2.6 DNA ligation of a forward adapter: mission impossible

We initially wanted to keep the 5'-end of the RNA molecules essentially random, with the exception of a terminal 5' rGrG, required for T7 RNA synthesis (**Figure 2-1**). The design made it necessary to ligate a second adapter after reverse transcription to allow a final PCR. In any case, a T7 promoter must be added to allow the next round of *in vitro* transcription. We have tested several DNA duplexes (we call it DNA anchor) with the aim to achieve DNA ligation with the 3'-end of the cDNA and also prevent PCR side products, especially primer dimers. These DNA anchors can be roughly categorized into 4 classes (**Figure 2-27**). The initial design (class I) is a double stranded T7 promoter oligo containing a sticky 3'-"GG" overhang intended to capture the 3'-"CC" end of the selected cDNA target. Classes II and III are identical to class I except for an extended sticky end region composed of a few additional constant positions following the GG overhang and either random bases (Class II) or a poly-Inosine (Class III). The role of these extensions is to help bind the selected cDNA and thus allow a better yield of ligation. We also tried a single stranded version of Class III (Class IV).

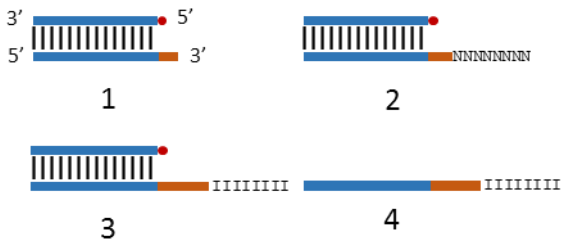



**Figure 2-27**. Four classes of DNA ligation anchors. These constructs are used to allow the ligation of the cDNA to a 5' adapter (T7 promoter) with a DNA ligase. Class I is a classic double stranded T7 promoter sequence with a "GG" sticky end. Class II is similar with class I but has a stretch of complementary and random bases following the "GG" sticky end. Class III has inosines instead of random bases. Class IV is a single stranded DNA oligo constituted by T7 promoter sequence, a cDNA overhang sequence and a poly-inosine tail. The T7 promoter sequence is in dark blue. The red dot stands for the 5'-end monophosphate group; The orange part indicates a sequence complementary to the targeted cDNA 3'-end.

All our constructs were tested for DNA ligation with a T4 DNA ligase (ThermoFisher). The product(s) of ligation were amplified by PCR and analyze on agarose gel.

No substantial PCR product could be obtained with our initial construct (Class I), even with overnight ligation at low temperature. Further attempts with our class II and III DNA anchors

showed that the yield of ligation could be improved, however at the expense of specificity: it was not possible to prevent the appearance of side-product during the final PCR (data not shown).

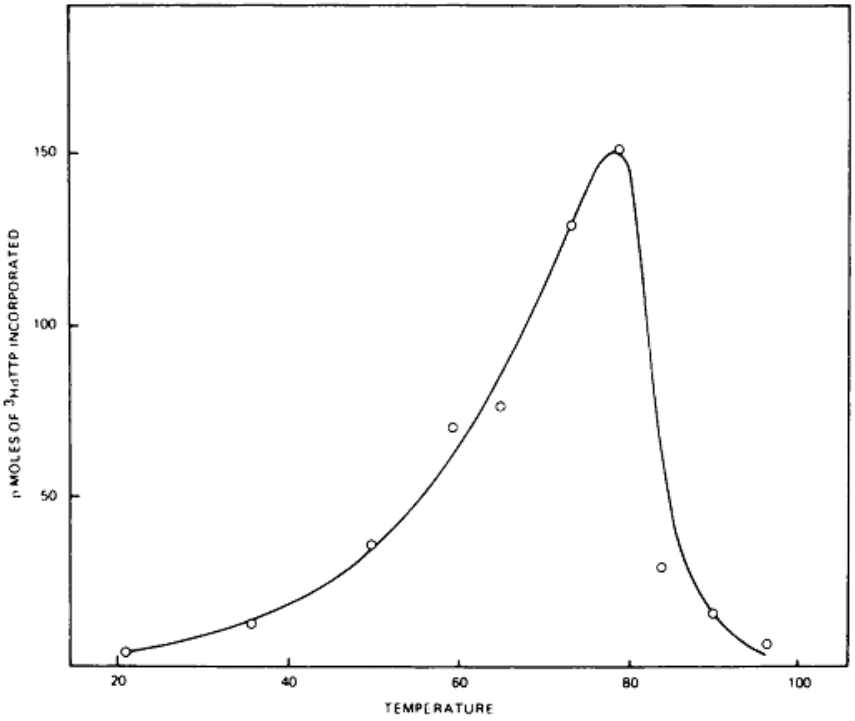



**Figure 2-28**: the correlation between temperature and relative activity of *Taq* DNA polymerase (Chien et al. 1976).

We also tried another approach: instead of relying on a DNA ligase to attach the forward T7 promoter, we tried to implement the addition of Class IV anchor during PCR with a specific annealing step. This procedure was taking advantage of the fact that a *Taq* DNA polymerase (ThermoFisher) still has some activity at low temperature (**Figure 2-28**).

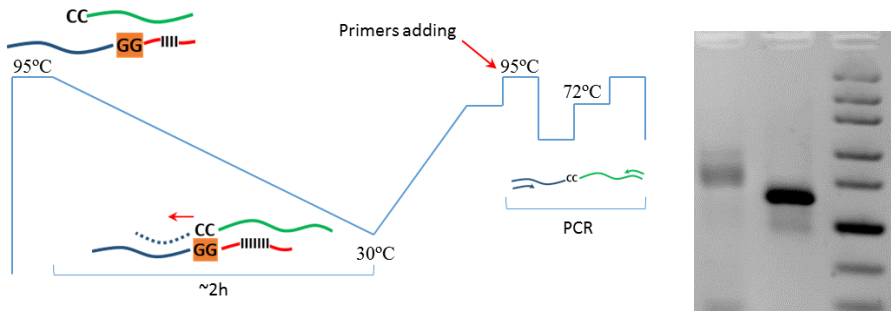



**Figure 2-29**. T7 promoter adding by *Taq* DNA polymerase: the mechanism of annealing-elongation procedure. Green: reverse transcribed target cDNA (3'→5'); Dark blue: T7 promoter (5'→3'); Dark blue dash line: elongated cT7 promoter sequence; Red line: poly-inosine tail (5'→3').

A pioneering research on *Taq* DNA polymerase (Chien et al. 1976) has shown that at temperatures below the optimum, like 40 to 60ºC, this polymerase still has a considerable activity. This is the reason why researchers prefer to use hot-start DNA polymerases for high-fidelity PCR. Our annealing-elongation method includes two main events (**Figure 2-29**): 1) Single strand DNA anchor and target cDNA are denatured at 95ºC and gradually cooled down to 30ºC in 2 hours, which allows the two oligos to anneal by both specific base-paring and inosine wobble base-paring. Notably, the 3'-"CC" end of target cDNA could only base pair with the opposite T7 promoter strand 3'-"GG" end, while the design of Adapter P doesn't allow the in excessed primer of RNA adapter to bind this DNA anchor. 2) Meanwhile, the *Taq* DNA polymerase was expected to slowly elongate the 3'-end of target cDNA using DNA anchor as template. After 2 hours of incubation, a reverse primer was added during denaturing, and a regular PCR program would go on from there. This strategy worked with a mini-tRNA (**Figure 2-29**), however with very low efficiency. Because we always experienced lower efficiencies with random pools, this approach was not pursued. In the end, we abandoned the idea of working with RNA pool completely randomized at the 5'-end. Instead, we decided to compromise and designed the shortest possible constant track the 5'-end that would allow direct PCR after the reverse transcription step (**Section 2.7**).

### 2.6.1 Reverse transcription

Reverse transcription (RT) is a reaction efficiency of which has been improved by engineered enzymes such as SuperScript® III (SSIII). With our initial design requiring forward adaptor ligation (see above), we however identified a critical issue related to terminal transferase activity and template-switching activity, known to occur with RT enzymes derived from Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) (Chen and Patton 2001), SSIII belongs to this family of enzymes. It was already suspected to have this behavior, and add extra non-templated nucleotides at the cDNA's 3'-end (Kulpa et al. 1997; Oz-Gleenberg et al. 2011, 2012; Zajac et al. 2013).

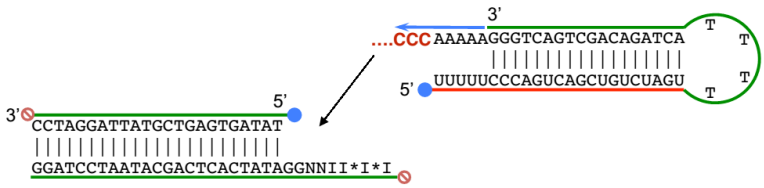



**Figure 2-30**. Primer-dimer generated by reverse transcription. Reverse transcriptase, like M-MLV RT, has a non-templated nucleotide addition activity. It preferably adds G or C at the cDNA's 3'-end. 3' Adaptors with added cytidines become compatible with the DNA anchor and undergo DNA ligation to generate a primer dimer. The green line represents the DNA oligo; The red line represents RNA oligo; Blue balls indicate the monophosphate groups; Red stop symbolizes an oligo modification preventing ligation.

Among the four deoxyribonucleotides, the MLV reverse transcriptase family prefers to add guanosine (G) and cytidine (C) at the cDNA's 3'-end. In our 5' DNA ligation attempts, all the tested DNA anchors necessarily had at least a "GG" sticky end (**Figure 2-27**). Thus, even though the 3' Adapter was carefully design to prevent Adapter/Anchor ligation, the non-template activity of SSIII made this adaptor compatible for DNA ligation, which generated a large amount of adapter/anchor dimer side-products. (**Figure 2-30**). This was the "last straw" that occurred in our DNA ligation attempts.

Before we dropped the DNA ligation strategy, we still tried to find a way to overcome the issue of SSIII non-templated activity. The Mung bean blunt enzyme is a nuclease derived from sprouts of mung bean *Vigna radiata* that degrades single stranded DNA or RNA to 5'-monophosphate (**Figure 2-31**). As a single-strand specific nuclease, it will not digest any double stranded DNA, double stranded RNA, or DNA/RNA hybrid oligos (McCutchan et al. 1984; Yu et al. 2014) provided the reaction occurs at a temperature below a certain threshold.



**Figure 2-31**. The Mung bean blunt enzyme only digests the single stranded RNA and DNA. Red color represents the RNA oligos; Green color indicates the DNA oligos; Dash line shows the digested single stranded RNA or DNA.

In order to test whether this enzyme could remove the non-templated additions of SSIII (**Figure 2-30**), a random RNA pool N37 (**Table 2-5**) was used as a template model to undergo RNA ligation, reverse transcription, Mung bean treatment, DNA ligation (**Figure 2-27**) and final PCR. **Figure 2-32** shows that Mung bean treated samples (lanes 3 and 4) can be rescued and give a clean band (100 pb), but only when this treatment is combined with the use of primers with overhangs during PCR (lane 4). Although the result suggested that this strategy could be

feasible, we anticipated from previous attempts that the extra steps required by Mung bean treatment and DNA ligation would be incompatible with the extremely low amount of RNA recovered after periodate oxydation. We therefore decided to abandon this strategy.

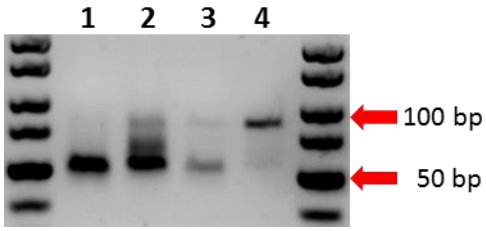



**Figure 2-32**. Random RNA pool N37 (37 nt) ligated with adapter P and transcribed into cDNA with SSIII RT. A purified cDNA sample was divided into two groups, with (3 and 4) or without (1 and 2) Mung bean treatment before DNA ligation. Lanes 1 and 3 show the PCR products obtained with T7 promoter primer (25 nt) & primer of the adapter (27 nt), while lanes 2 and 4 show the result of PCR using T7 promoter+GGCACGA overhang (30 nt) & primer of the adapter+TAG overhang (30 nt). Only the sample (lane 4) treated by blunt enzyme and amplified by overhanged primers can generate the expected PCR product without any side-product.

## 2.7 PCR optimization

A successful PCR amplification reveals if all steps in a selection cycle go well, in which case a pool can be regenerated for the next selection cycle. Considering the difficulties encountered with (almost entire) random RNA pools, that require the ligation of a 5' anchor, we decided to abandon this strategy. Instead, we designed the shortest possible constant track at the 5' end of these RNA that could allow the overhang of a T7 promoter to prime the corresponding cDNA during final PCR. The only difficulty is to be able to perform an efficient amplification while using the extremely short conserved region. An optimal PCR requires a melting temperature above 45ºC and below 72ºC in a classic thermocycler reaction (Hyndman and Mitsuhashi 2003). Previous investigations have shown that short primers around 10 nt can achieve reliable amplification. Furthermore, it has been reported that chemical compounds like the tripeptide 1,2-dihydro-(*3H*)-pyrrolo-[3,2-*e*] indole-7-carboxylate ($CDPI_3$), can bind to the minor groove of DNA with high affinity and allow primers as short as 8 to 10-mers to perform specific and efficient amplification (Afonina et al. 1997). GC rich primer as short as 7-mers can be used in PCR by simply reducing the annealing temperature down to $35^{o}$ C with however much less specificity (Vincent et al. 1991). Minimal primer with only 6-mers can also be used to

amplify the plasmid carrying some specific restriction sites (Ryu et al. 2000).

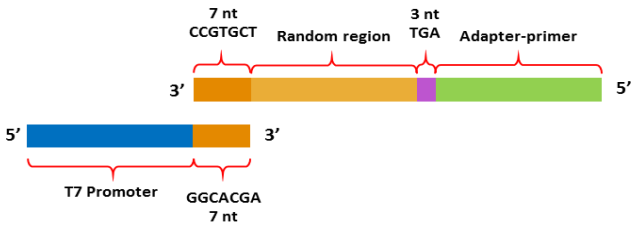



**Figure 2-33**. cDNA structure and T7 promoter overhang primer design. The upper strand briefly shows the structure of the cDNA. The green part represents the added adapter part; The purple and orange parts indicate the conserved sequence of the initial RNA template; The light yellow part shows the random track of the initial RNA template. The bottom strand shows the structure of the T7 promoter with overhang primer. The blue part represents the T7 promoter sequence and the orange part indicates the overhang sequence priming on the upper cDNA orange part.

Based on these data, we designed a series of RNA pools characterized by 7 constant positions at the 5' end and 3 specified nucleotides at the 3' end. A forward primer carrying a T7 promoter with a 7-nt overhang was designed to prime with the cDNA's 3'-end as shown in **Figure 2-33**. Obviously, only 7 base pairs are not sufficient enough to provide strong binding during the annealing step. However, the predicted melting temperature is around 25° C in a standard PCR mixture with 2 mM $MgCl_2$, implying that we could take advantage of the low activity of *Taq* DNA polymerase at room temperature (see **Section 2.6**). This design was initially tested with a ssDNA similar to the considered cDNA. T7 promoters with three different overhang (7, 8 and 9 nt) were tested using both a *Taq* DNA polymerase (ThermoFisher) and a *Taq* Hot Start DNA polymerase (NEB). When the annealing temperature was set at 60° C, only the T7 promoter with a 9-nt overhang could produce efficient amplification with the Hot Start *Taq* polymerase (**Figure 2-34**, 3'), while the 7 nt overhang T7 promoter only generate a very faint signal. When the PCR mixture is prepared at room temperature with a *Taq* DNA polymerase, the final PCR can yield strong and specific signal with 7-, 8-, and 9-nt overhang, indicating that the considered (GC-rich) 7 nt overhang is long enough to prime and PCR the cDNA. Remarkably, no unspecific product is generated (**Figure 2-34**). This experiment demonstrated that RNA pools with as little as 7 constant positions at the 5' end and 3 constant positions at the 3' could undergo specific PCR amplification after selection.

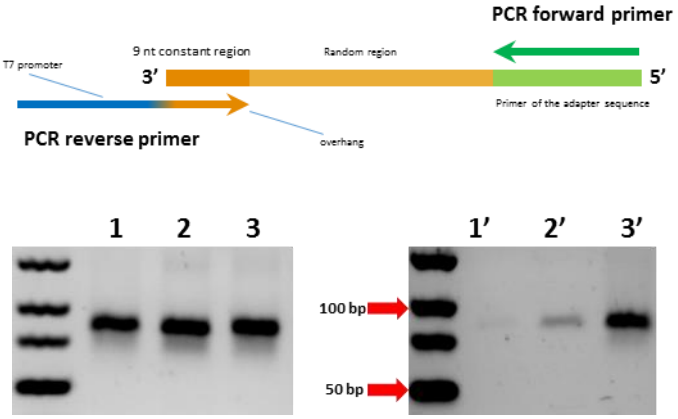



**Figure 2-34**. Short primers PCR procedure test. Left: PCR products with DreamTaq DNA polymerase (ThermoFisher) using T7 primers with 7 nt (1), 8 nt (2) and 9 nt (3) overhangs together with a forward primer. Right: PCR product with OneStart *Taq* DNA polymerase (NEB) using T7 primers with 7 nt (1'), 8 nt (2') and 9 nt (3') overhangs together with a forward primer.

## 2.8 The final SELEX strategy

All the refinements elaborated throughout the assessment of each step of our SELEX strategy could finally be combined into a working protocol (**Figure 2-36**). Compared with the initial design (**Figure 2-1**), the major change implemented in the final design is the absence of a DNA ligation step after reverse transcription. This step was investigated for almost two years, and proved to be impossible. It could be remove at the expense of the presence of a 7-nt constant track at the 5' end of the RNA pools, implying that these pools would still impose some constraints on the kind of ribozyme that could be isolated.

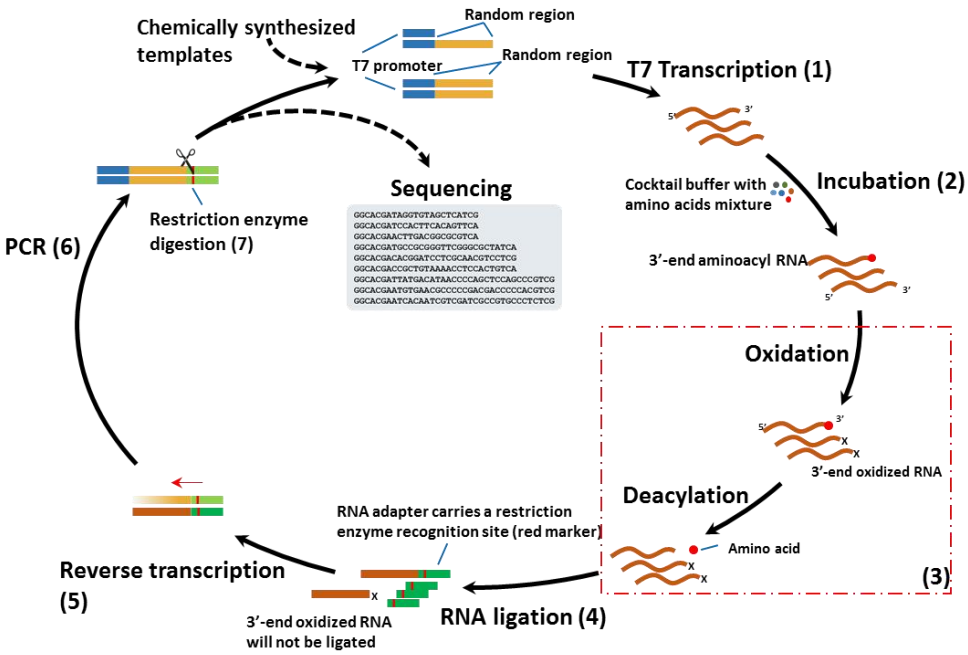



**Figure 2-35**. The final SELEX protocol. 1) Chemically synthesized DNA templates or treated PCR products which include very short conserved bases are used to transcribed into a pool of random RNA; 2) Pool of RNA incubates with complex buffer; 3) After incubation, RNA is treated by periodate. The 3'-end non-aminoacylated RNA is oxidized. Followed by deacylation, the amino acid is released from the 3'-end of aminoacyl RNA and make the 3'-end available for adapter ligation; 4) 3'-end non-oxidized RNA ligates with a carefully designed RNA adapter carrying a restriction enzyme recognition site by T4 RNA ligase; 5) Ligated RNA is reverse transcribed to cDNA; 6) 7nt overhang short primer PCR to add the T7 promoter sequence; 7) Restriction enzyme treatment removes the adapter part and generates the new DNA templates for next round selection.

# 3. Results and discussion

## 3.1 SELEX experiment

Altogether, four different RNA libraries corresponding to the optimal design of the pools (**Table 2-5**) went through 4 to 7 cycles of selection (**Figure 2-36**), during which they were incubated with a mixture of 19 different kinds of free amino acids with ATP and GTP as activating cofactors. In order to quickly proceed through the cycles, only the **OD** treatment was achieved during the selection step; the **DO, D** and **NT** controls (**Section 2.4.2**) were done only at the end, after a convergence was detected. The initial selection started with a pool of RNA including approximate $10^{13}$ RNA molecules in each library. (**Table 3-1**).

**Table 3-1**. Sequence diversity of initial pool

| Name | Diversity | Initial amount | Copy number of each unique sequence |
|---|---|---|---|
| N19 | $2.6\times10^{5}$ | 0.025 nmol | $\sim10^{8}$ |
| N23 | $6.7\times10^{7}$ | 0.025 nmol | $\sim10^{6}$ |
| N30 | $1.1\times10^{12}$ | 0.025 nmol | ~10 |
| N37 | $1.5\times10^{13}$ | 0.025 nmol | 0 or 1 |

After each incubation reaction, all four samples underwent the same **OD** treatments and were amplified by RT-PCR. PCR amplification after the **1**$^{st}$ cycle of selection was not easy to achieve due to the very low amount of remaining (non-oxidized) RNA after **OD** treatment. The usual amount of solution from the reverse transcription reaction (2 $\mu$l out of 20 $\mu$l in total) had to undergo as many as 35 cycles of PCR to only generate a faint signal on gel. In total, 5 tubes of 50 $\mu$l-PCR reaction were combined for gel purification.

The quality and diversity of the pools was verified after the first round selection. The PCR products were used in a cloning experiment using the CloneJET kit (ThermoFisher), and about 10 individual clones in each pool were sent for sequencing. The sequencing results for each pool are shown in **Table 3-2, A-D** (the red letters highlight the constant regions). As expected at this stage, the 35 sequences in total are all different and do not show any similarity either in sequence or in their predicted secondary structure (established using *Mfold*; Zuker 2003).

**Table 3-2, A**. Sequencing results of N19 pool after **1**st cycle selection

| Sequence | Optimal secondary structure |
|---|---|
| **GGCACGA**AGAAAAUUG**UCA** | ((((.....)))). |
| **GGCACGA**AUACCCGCC**UCA** | (((.....)))... |
| **GGCACGA**ACUUCCACG**UCA** | ...(((.....))).. |
| **GGCACGA**UGGCCGUGU**UCA** | .(((((.....))))).. |
| **GGCACGA**ACUUAUCCC**UCA** | ..... |
| **GGCACGA**CUACCACAC**UCA** | ..... |
| **GGCACGA**UGUACCUUUG**UCA** | ((((.....)))). |
| **GGCACGA**CUGCCACAACAAU**UCA** | ((((....)))).... |
| **GGCACGA**GCCGGCAAC**UCA** | (((....)))..... |
| **GGCACGA**ACAACACAG**UCA** | ..... |

**Table 3-2, B**. Sequencing results of N23 pool after **1**st cycle selection

| Sequence | Optimal secondary structure |
|---|---|
| **GGCACGA**CCACAUUACGGCC**UCA** | (((.((.....)))))... |
| **GGCACGA**GCCCACCGCGAGA**UCA** | (((....)))..... |
| **GGCACGA**AACGACUCUGCCU**UCA** | ((((.((.....)))))).... |
| **GGCACGA**AUUACAUUUAGACUUUU**UCA** | ..... |
| **GGCACGA**CACAAUGCCGUCC**UCA** | ((((.....)))).... |
| **GGCACGA**UUCUACCAAUAGU**UCA** | ..... |

**Table 3-2, C**. Sequencing results of N30 pool after **1**st cycle selection

| Sequence | Optimal secondary structure |
|---|---|
| **GGCACGA**AGGAGCUUGCGUAGUG**UCA** | ((((((...(.....)...)))))). |
| **GGCACGA**UCAUCCCCUCUUGAAUA**UCA** | .....(((.....))).... |
| **GGCACGA**AUCCUAACCCAAGUCCCCCU**UCA** | ((.((.....)).))..... |
| **GGCACGA**GUCGAUCUUUCCGCCAAUUG**UCA** | (((..(((.....)))...))).... |
| **GGCACGA**CAUGGCCCUCCCCAGAAUAU**UCA** | (((.....))).... |
| **GGCACGA**UGCCGCGGGUUCGGGCGCUA**UCA** | ((((...))))(((.....))).... |
| **GGCACGA**ACAAACAUGUUACGCGCUCC**UCA** | (((.((((((....)))).)).)))..... |
| **GGCACGA**GGGCACGAACAAGUUAUUGC**UCA** | .....(((((.....))))). |
| **GGCACGA**CCGUAUUGUGAGGCCGUGUGUGA**UCA** | .((((((.(.....)).)))))..... |

**Table 3-2, D**. Sequencing results of N37 pool after **1**st cycle selection

| Sequence | Optimal secondary structure |
|---|---|
| **GGCACGA**UAUCGUCGCCUCC**UCA** | ((((((....))).)))..... |
| **GGCACGA**ACCCAACGCGGUCUCGCCC**UCA** | (((..((.((.....))))..)))).... |
| **GGCACGA**UGUUGAAGUACUAGUACAUCCCGCC**UCA** | (((..(((((.....))))))..)))... |
| **GGCACGA**ACACACAAUUCACCGGACUUCCAAGUG**UCA** | ((((((.....)))))). |
| **GGCACGA**CCCAUUGCACAUUAACGCUCACCCUAG**UCA** | .(((.....)))..... |
| **GGCACGA**CAACUACUCGCCCCGUGAGUCUACAUG**UCA** | .....((((...((((((...))))))....)))). |

| | |
|---|---|
| **GGCACGA**CUCCUACACCAUAGAAGCGUUAAGCG**UCA** | .((((((.((.(((.....))).))))))...)).... |
| **GGCACGA**AACCACACCAAACUUGCAGUAACCAAA**UCA** | .(((.....))).... |
| **GGCACGA**CGAUCAACUACAGCUUGCUGCGCCGCA**UCA** | (((.((.(((.(.....).))).)).))).... |
| **GGCACGA**CGCAUCGAUGCCCGUCGCUCUCACAAG**UCA** | (((((((....))).)))).... |

It can however be noted that despite the use of reverse primers with TGA 3' overhangs (specific to the 3' ends of the pools) the size of the clones already experienced some fluctuations after the first cycle.

The obtained PCR products were re-amplified using the AccuPrime™ *Taq* polymerase (Invitrogen), with which very clean products could be obtained (**Figure 3-1**, left). The PCR products were cleaned up and digested with the EarI restriction endonuclease (NEB) (**Figure 3-1,** right).

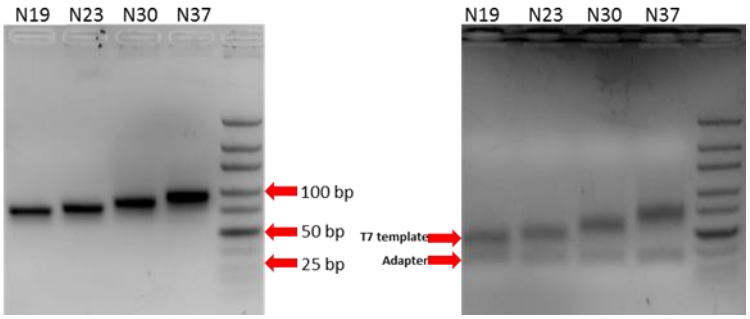



**Figure 3-1**. Left: After the first selection cycle, 1 $\mu$g of PCR products were loaded on 4% agarose gel. Right: The same PCR products (1 $\mu$g) after digestion by the EarI restriction enzyme at 37ºC overnight.

The digested PCR products were directly used for T7 transcription without further purification. Since the EarI digestion could not reach completion, a small fraction of non-digested PCR products carrying the RNA adapter sequence were still present, and could be observed during the next round RNA synthesis (**Figure 3-2**). This long RNA transcript did not seemingly affect the next rounds of the selection experiment.

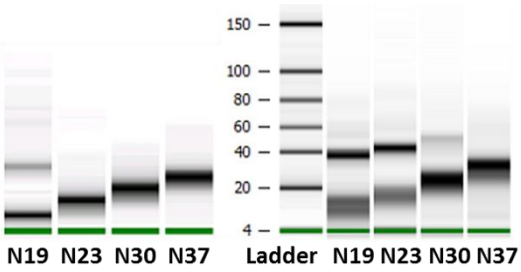



**Figure 3-2**. Bioanalyzer signals of initial RNA pool (left) and RNA pool after the first selection cycle (Right). A longer transcript is commonly observed from the second round.

Following the experimental strategy summarized in **Figure 2-36**, all four RNA libraries went through successive **OD** treatment and re-amplification for the remaining cycles. From the **2**nd cycle, cDNA libraries became much easier to regenerate and less PCR cycles were required after the reverse transcription step. After three selection cycles, we noticed that our standard DNase treatment procedure could not eliminate all DNA after T7 transcription. Since this could critically impair the selection process, we decided to use HPLC to purify the RNA transcripts for the remaining cycles (see **Supporting information** HPLC purification). Since a SELEX procedure reduces the sequence diversity after each cycle, the amount of RNA used in the experiments was decreased from 0.025 nmol (cycles 1 to 3) to 0.01 nmol (from cycle 4). The related parameters are indicated in **Table 3-3**.

**Table 3-3**. SELEX parameters during the selection cycles

| | RNA Purification | | Initial quantity (nmol) | | RT-PCR cycles | |
|---|---|---|---|---|---|---|
| Library name | Cycle 1-3 | Cycle 4-6 or 7 | Cycle 1-3 | Cycle 4-6 or 7 | Cycle 1 | Cycle 4-6 or 7 |
| N19 | Gel | HPLC | 0.025 | 0.01 | 35 (50 $\mu$l x 5) | 30 (50 $\mu$l x 2) |
| N23 | Gel | N/A | 0.025 | N/A | 35 (50 $\mu$l x 5) | N/A |
| N30 | Gel | HPLC | 0.025 | 0.01 | 35 (50 $\mu$l x 5) | 30 (50 $\mu$l x 2) |
| N37 | Gel | HPLC | 0.025 | 0.01 | 35 (50 $\mu$l x 5) | 30 (50 $\mu$l x 2) |

After only four cycles of selection, the N23 pool became highly homogenous. This phenomenon was noticed during a bioanalyzer analysis (**Figure 3-3**): unlike specific RNA molecules, a random RNA pool normally shows a much boarder peak on the bioanalyzer panel due to the distribution of the secondary structures and length heterogeneity. We decided to pause the selection for this pool and to continue working with the other libraries. **Figure 3-3** shows the bioanalyzer traces of the initial pools and pools of RNA observed at the final cycle. We

observed that the peaks of all libraries (except for N23) became much sharper after 6 or 7 rounds of selection. This behavior suggested to us that the selection process had converged. Consequently, we decided to proceed with sequence analysis rather than to continue with more cycles.

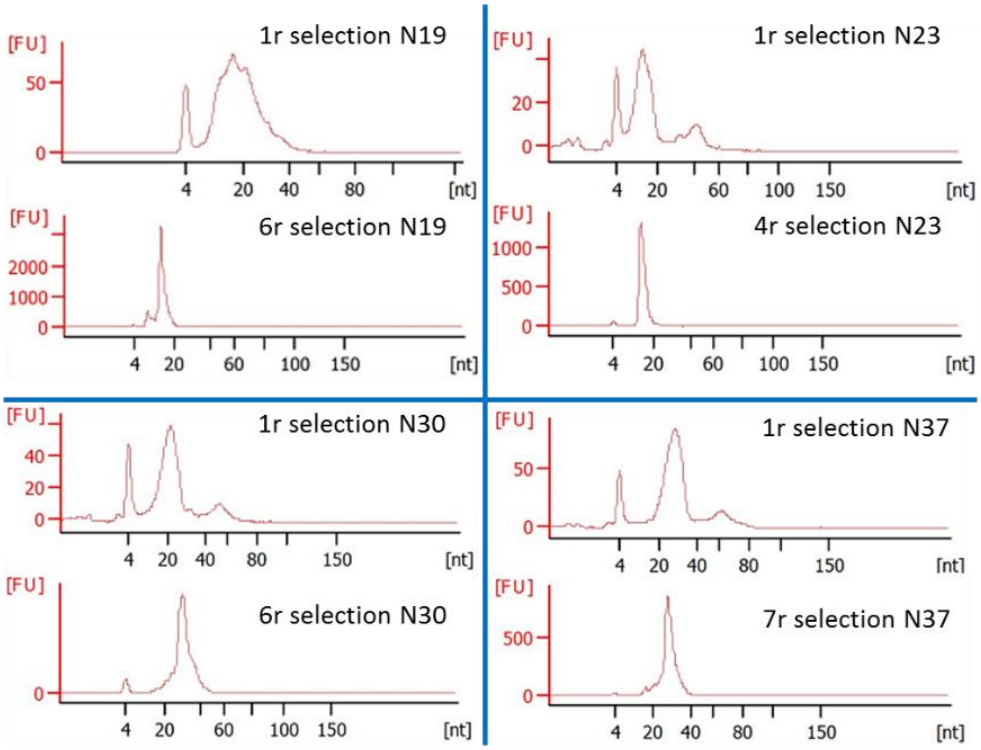



**Figure 3-3**. Bioanalyzer traces of the initial pools and the final pools. The N23 library went through **4** cycles of selection; the N19 and N30 library were run until they reached the **6**th selection cycle; the N37 pool was further pushed to the **7**th cycle. Note that the scale of the x-axis is not uniform.

## 3.2 Sequence analysis after the final round of selection

After the final cycle, the PCR products of each four libraries were inserted into pJET plasmids using the CloneJET PCR Cloning kit. Single clones were sent for sequencing, and a total number 156 sequences were obtained. We first discuss the results obtained with each of the four libraries. An overall analysis in which these pools are compared to each other is presented afterwards.

### Library N19

After 6 rounds of selection, 46 independent sequences were collected from library N19 (**Table 3-4**). The length distribution is consistent with the expected size: 50% of the sequences

still have the initial size of 19 nt, while the remaining sequences are slightly longer or shorter. It can already be noticed that an appreciable fraction has a size of 23 nt. In terms of diversity, only around 40% of the sequences are unique, which means the SELEX procedure eliminated about 60% of the diversity of the pool.

**Table 3-4**. Sequencing result of library N19, **6**th round of selection

| No. | Length | Sequence (5'→3') | Copy # | Secondary structure |
|---|---|---|---|---|
| 1 | 18nt | **GGCACGA**ACAGCAUG**UCA** | 1 | ((((.(.....).)))). |
| 2 | 19nt | **GGCACGA**ACCACAGUG**UCA** | 10 | (((((.........))))). |
| 3 | | **GGCACGA**CUAGCCAGC**UCA** | 2 | (((.......)))..... |
| 4 | | **GGCACGA**ACACUAGUG**UCA** | 1 | (((((.........))))). |
| 5 | | **GGCACGA**ACAAUGAUG**UCA** | 1 | (((((.......).))))). |
| 6 | | **GGCACGA**ACAUUGAUG**UCA** | 1 | (((((((...))).))))). |
| 7 | | **GGCACGA**ACACUCUUG**UCA** | 1 | ((((.((....)).)))). |
| 8 | | **GGCACGA**ACAUCGACG**UCA** | 1 | (((.(((...))))..))). |
| 9 | | **GGCACGA**ACAGCAACG**UCA** | 1 | .((.......))..... |
| 10 | | **GGCACGA**ACAGUUCCG**UCA** | 1 | ((.((.....)).))..... |
| 11 | | **GGCACGA**ACCACUGUG**UCA** | 1 | ((((((.......)))))). |
| 12 | | **GGCACGA**CUAUUCGAC**UCA** | 1 | ....(((....)))..... |
| 13 | | **GGCACGA**CGAACGCCA**UCA** | 1 | (((.........)))..... |
| 14 | | **GGCACGA**CAGCACGCC**UCA** | 1 | (((.........)))... |
| 15 | 21nt | **GGCACGA**CUAGACAGCCAA**CA** | 3 | (((.........)))..... |
| 16 | | **GGCACGA**CUAGACAGCCA**UCA** | 1 | (((.........)))..... |
| 17 | 23nt | **GGCACGA**CUAGACAGCCAAC**UCA** | 12 | (((.........)))..... |
| 18 | | **GGCACGA**CUAGCCAGCCCCG**UCA** | 1 | (((.......)))..... |
| 19 | | **GGCACGA**UAGCCAGCACUGC**UCA** | 1 | (((.....)))(((...)))).. |
| 20 | | **GGCACGA**ACAGCCAUCCGAC**UCA** | 1 | (((.......)))..... |
| 21 | | **GGCACGA**CGUACGAUCCAUG**UCA** | 1 | .....(((((.......))))). |
| 22 | | **GGCACGA**AGCUACUCGUCGG**UCA** | 1 | (((((((.....))))..))). |
| 23 | 26nt | **GGCACGA**UUAUUAGUCAACUAAC**UCA** | 1 | (((.........)))..... |

An examination of these sequences reveals that they share some similarities (**Table 3-4**). In particular, many of them bear a same ACA motif that follows the 5' GGCACGA constant track (**Table 3-4**, underlined). This motif is most often found inside a loop in the predicted secondary structures, suggesting that it could have a structural or functional role. These secondary structures also exhibit some similarities. Most of them fold into a stem-loop with a single-stranded 3' end. However, many of these free 3' ends only consist of the terminal "A" residue that is not expected to be flexible enough to reach any potential catalytic site (see

Lehmann et al. 2007). It does suggest that this kind of species is probably a false positive result. The secondary structure of these RNA may assist them to avoid periodate oxidation and/or allow a much more efficient RNA ligation.

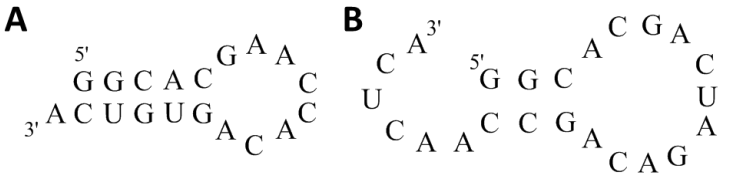



**Figure 3-4**. Predicted secondary structure (Mfold) of the most enriched sequences in N19 library. **A**: the "*Key*" shape secondary structure of No.2 (10 copies); **B**: the "*Hammer Throw*" shape secondary structure of No.17 (12 copies).

For the longest selected RNA, the free single-stranded 3'-end is usually much longer (**Figure 3-4**, B), a structure that would be consistent with a catalytic activity: similar to the tiny Yarus ribozyme (Illangasekare and Yarus 1999; Lehmann et al. 2007), the free 3'-end could reach a plausible binding pocket constituted by the large predicted loop. This "*Hammer Throw*" shape secondary structure is commonly found among those sequences; it includes a 7 to 11-nt loop, a three base-pair stem and a single stranded 3'-end tail of different lengths.

**Library N23**

The selection of library N23 was stopped after the **4**th round because the bioanalyzer trace showed a very sharp peak (**Figure 3-3**), that we initially suspected to be due to a sequence contamination, or an accident that created a massive loss of sequence diversity. A small sample was sent for sequencing, from which 13 independent sequences were obtained (**Table 3-5**). Interestingly, 11 out of 13 sequences still keep the correct size of 23 nt. These sequences are almost all different, but 2 out of 13 sequences (**Table 3-5**, No.1) are identical to 2 sequences found in library N19. Since all the experiments were performed independently, we are very confident that is not a cross contamination. In terms of secondary structure, no clear motif could be detected besides the common stem-loop structure (with either a *Key*" or "*Hammer Throw*" shape). Two sequences (No. 5 and 9) have no predicted secondary structure, suggesting that the pool was not yet converging; it seems we gave up with library N23 a bit too early.

**Table 3-5**. Sequencing result of library N23, **4**th round of selection

| No. | Length | Sequence | Copy # | Secondary structure |
|---|---|---|---|---|
| 1 | 23nt | **GGCACGA**CUAGACAGCCAAC**UCA** | 2 | (((..........)))..... |
| 2 | | **GGCACGA**AGCCAGCUCCCCG**UCA** | 1 | (((.....)))..... |
| 3 | | **GGCACGA**UGUGGACUGCAGC**UCA** | 1 | .....((..((.....))..)). |
| 4 | | **GGCACGA**AGCAUCGAACAGC**UCA** | 1 | .....(((.....)))..  |
| 5 | | **GGCACGA**UCCUUAACCAUAC**UCA** | 1 | ..... |
| 6 | | **GGCACGA**UCACCCUGUGAUG**UCA** | 1 | (((((...((((...)))))))). |
| 7 | | **GGCACGA**CUAGUCAGCGUCG**UCA** | 1 | ...(((((.....)))))..  |
| 8 | | **GGCACGA**CCCUGCGCGGGGC**UCA** | 1 | .....((((((....)))).)). |
| 9 | | **GGCACGA**ACCCCCUUACGAC**UCA** | 1 | ..... |
| 10 | | **GGCACGA**CUACCCCCACCGC**UCA** | 1 | (((.....)))..  |
| 11 | 24nt | **GGCACGA**CCGUCUCCCAACAC**UCA** | 1 | (((.....)))..... |
| 12 | | **GGCACGA**CCACGUGACCCCCC**UCA** | 1 | (((((((....)))).)).....  |

### Library N30

It is very instructive to already compare the sequencing result of library N30 (**Table 3-6**) with the two previous libraries. The length distribution is quite board, and no clone kept the initial size of 30 nt. However, among the 24 identified sequences, many of them are absolutely identical to sequences found in libraries N19 and N23, revealing that a same convergence phenomenon occurred independently, during which not only the sequence but also the size underwent a selection process.

**Table 3-6**. Sequencing result of library N30, 6th round of selection

| No. | Length | Sequence | Copy # | Secondary structure |
|---|---|---|---|---|
| 1 | 19nt | **GGCACGA**CAGCACGCC**UCA** | 2 | (((.....)))... |
| 2 | 20nt | **GGCACGA**CAGCCCCGCC**UCA** | 2 | (((.....)))..... |
| 3 | | **GGCACGA**ACAGCACGAC**UCA** | 1 | .((.....))..... |
| 4 | 22nt | **GGCACGA**ACAUCUCCCGGC**UCA** | 1 | (((.((.....)))))..  |
| 5 | 23nt | **GGCACGA**CAGCUACCCGCCC**UCA** | 1 | (((.....)))..... |
| 6 | | **GGCACGA**GACCGCUCCCGAC**UCA** | 1 | ((...(((....)))))..... |
| 7 | | **GGCACGA**GACAGCACCCGCC**UCA** | 1 | (((.....)))... |
| 8 | | **GGCACGA**GACACCUACCGCC**UCA** | 1 | (((...((....))...)))... |
| 9 | | **GGCACGA**CUAGACAGCCAAC**UCA** | 1 | (((.....)))..... |
| 10 | | **GGCACGA**CAGCCACCGCCCC**UCA** | 1 | (((.....)))..... |
| 11 | | **GGCACGA**CUCUGCGAGGCCC**UCA** | 1 | (((.((.....))..)))....  |
| 12 | 26nt | **GGCACGA**CUAGUGCAGAGCAACC**UCA** | 1 | .((((.....)))).(((....))). |

| | | | | |
|---|---|---|---|---|
| 13 | 34nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 1 | ((.(((.....))).))....(((.....))). |
| 14 | 36nt | **GGCACGA**CCAGCGUUCCUCUAGACCCCUACGUC**UCA** | 4 | ((.(((.....))).))...((((.....)))).. |
| 15 | 37nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUCU**UCA** | 5 | ((.(((.....))).))...((((.....)))).... |

In addition, the longest sequences of library N30 are also identical to the most abundant clones identifies in the N37 library (see below).

**Table 3-7**. Sequencing result of library N37, **6**th round of selection

| No. | Length | Sequence | Copy # | Secondary structure |
|---|---|---|---|---|
| 1 | 23nt | **GGCACGA**CUAGACAGCCAAC**UCA** | 1 | (((.........))).... |
| 2 | 30nt | **GGCACGA**ACAAACAUGUUACGCGCUCC**UCA** | 1 | (((.(((((((....)))).)).))).... |
| 3 | | **GGCACGA**ACCACGCUACCUGACCCACG**UCA** | 1 | (((.........)))...((((....)))) |
| 4 | 34nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 2 | ((.(((.....))).))....(((.....))). |
| 5 | 36nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUC**UCA** | 1 | ((.(((.....))).))...((((.....)))).. |
| 6 | 37nt | **GGCACGA**CCAGCGUUCCCCUGGACCCCUACGUCU**UCA** | 1 | ((.(((.....))).))...((((.....)))).... |
| 7 | | **GGCACGA**CCGCCAUCGCCACUCCAACGUGGUCAC**UCA** | 1 | (((.....)))...(((((.....))))).... |
| 8 | | **GGCACGA**CCACUCCCGUGGAACCACAGCAACGCC**UCA** | 1 | .....((((....))))..... |

**Table 3-8**. Sequencing result of library N37, **7**th round of selection

| No. | Length | Sequence | Copy # | Secondary structure |
|---|---|---|---|---|
| 1 | 15nt | **GGCACGA**UCAUG**UCA** | 1 | ((((.....)))). |
| 2 | 18nt | **GGCACGA**CCCGCGCC**UCA** | 1 | (((.((...)).)))... |
| 3 | | **GGCACGA**CCCACGCC**UCA** | 1 | (((.........)))... |
| 4 | 19nt | **GGCACGA**CAUCUAACG**UCA** | 1 | .....(((.......))). |
| 5 | | **GGCACGA**CAGCACGCC**UCA** | 1 | (((.........)))... |
| 6 | 20nt | **GGCACGA**CAGCCCCGCC**UCA** | 1 | (((.....)))..... |
| 7 | 23nt | **GGCACGA**ACAGCUGCCCGUC**UCA** | 1 | (((((.....).)))).... |
| 8 | | **GGCACGA**CUAGCCAGCCAAC**UCA** | 1 | (((.......)))..... |
| 9 | | **GGCACGA**CUAGACAGCCAAC**UCA** | 11 | (((.........)))..... |
| 10 | | **GGCACGA**GACACCUACCGCC**UCA** | 7 | (((...((....))...)))... |
| 11 | | **GGCACGA**GACAUCUAUCGCC**UCA** | 1 | (((...((....))...)))... |
| 12 | | **GGCACGA**GACAUCCAACGCC**UCA** | 1 | (((..((....))....)))... |
| 13 | | **GGCACGA**GACACCCAACGACC**UCA** | 1 | .....(((.(.......).))). |
| 14 | | **GGCACGA**UAGCCAGCACUGC**UCA** | 1 | (((.....)))(((...))).. |
| 15 | | **GGCACGA**GACAGCUCCCAAC**UCA** | 1 | (((.......)))..... |
| 16 | 27nt | **GGCACGA**GACAUGCCACACCGCAC**UCA** | 1 | ((((.......))))..... |
| 17 | | **GGCACGA**CUAGACACCUCACGACC**UCA** | 3 | ..... |
| 18 | 28nt | **GGCACGA**CUAGACAUCUAACUCCCG**UCA** | 1 | .....(((..((.......))..))). |
| 19 | 29nt | **GGCACGA**CCAGCGUUCCUCUGGACCC**UCA** | 1 | .....((((.(...).))))..... |
| 20 | 30nt | **GGCACGA**ACUAGCCAGCAUUCGACUCG**UCA** | 2 | (((.......)))..... |
| 21 | 34nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 12 | ((.(((.....))).))....(((.....))). |
| 22 | 36nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUC**UCA** | 6 | ((.(((.....))).))...((((.....)))).. |
| 23 | 37nt | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUCU**UCA** | 7 | ((.(((.....))).))...((((.....)))).... |

**Library N37**

Since the N37 library is experimentally easier to handle, we run one more selection cycle with this pool. Altogether, 9 clones were sequenced from round **6** (**Table 3-7**), while 64 sequences were identified from round **7** (**Table 3-8**). From the table, we can see that the diversity of the pool has dropped dramatically comparing with the first selection cycle (**Table 3-2**): most sequences share some similarities in their secondary structures (e.g. **Table 3-8**, No.10 to 13), such as the one shown in **Figure 3-5**.

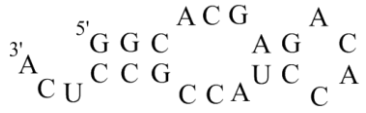



**Figure 3-5**. Optimal secondary structure (Mfold) of one of the enriched sequence No.10 (7 copies) in **Table 3-8**.

Half of the sequences can fold into a single loop with an unstructured 3'-end while the others can form two hairpin structures. The most abundant structural motif is present in sequences No. 21, 22 and 23 (**Table 3-8**), that are highly conserved with only a few point mutations close to the 3'-end. These differences can slightly change the 3'-end hairpin folding (**Figure 3-6**). This kind of structures share similarities with the Yarus C3 ribozyme (Chumachenko et al. 2009). It may indicate that a SELEX procedure lead to the selection of highly structured and stable RNA molecules.

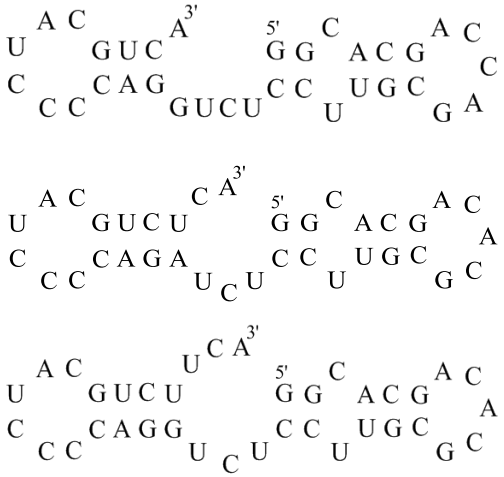



**Figure 3-6**. Optimal secondary structure of No. 21, 22 and 23 in **Table 3-8**. Their 5'-end hairpins are the same while point mutations present at the 3'-end can slightly change the size of stem as well as the length of unstructured 3'-end residues.

Considering altogether the three libraries N19, N30 and N37 at their final round, a striking

phenomenon occurred during selection: despite the conserved regions at both the 5' and 3' ends combined with the use of primers with corresponding overhangs, there was a propensity for all the three pools to generate 23-nt sequences (**Figure 3-7**). Thus, despite our strategy to keep the size constant, a strong selection mechanism forced all the pools to adopt the size and some typical sequences of the N23 pool.

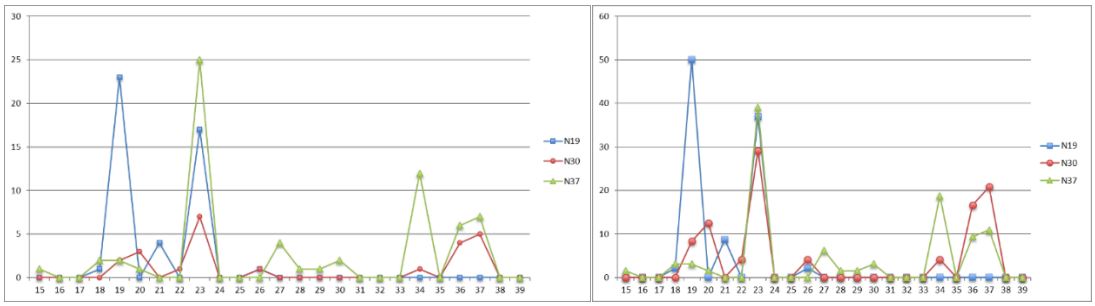



**Figure 3-7**. Left: Length distribution of libraries N19, N30 and N37. Right: Normalized length distribution for all three libraries. The X axis stands the length of sequences. The Y axis is the number of sequences.

Thus, it turned out that the "strange" behavior of the N23 pool, that seemed to prematurely converge towards certain sequences, was likely due to an "attractor effect" that also affected all three other pools.

A comparison of all libraries shows a convergence towards certain structures that occurred for all of them. A classification of these structures into seven categories is presented in **Tables 3-9** and **3-10**.

**Table 3-9**. Categories I to V

| Cat. | Sequence | Copy # | Secondary structure | Library |
|---|---|---|---|---|
| I | **GGCACGA**CUAGCCAGC**UCA** | 2 | (((.....)))..... | N19(6r) |
| | **GGCACGA**CUAGCCAGCCCCG**UCA** | 1 | (((.....)))..... | N19(6r) |
| | **GGCACGA**CUAGCCAGCCAAC**UCA** | 1 | (((.....)))..... | N37(7r) |
| II | **GGCACGA**ACAGCAACG**UCA** | 1 | .((.....))..... | N19(6r) |
| | **GGCACGA**ACAGCCAUCCGAC**UCA** | 1 | (((.....)))..... | N19(6r) |
| | **GGCACGA**ACAGCACGAC**UCA** | 1 | .((.....))..... | N30(6r) |
| III | **GGCACGA**CAGCACGCC**UCA** | 1 | (((.....)))... | N19(6r) |
| | **GGCACGA**CAGCACGCC**UCA** | 2 | (((.....)))... | N30(6r) |
| | **GGCACGA**CAGCACGCC**UCA** | 1 | (((.....)))... | N37(7r) |
| IV | **GGCACGA**CUAGACAGCCA**UCA** | 1 | (((.....))).... | N19(6r) |
| | **GGCACGA**CUAGACAGCCA**ACA** | 3 | (((.....))).... | N19(6r) |

| | | | | |
|---|---|---|---|---|
| | **GGCACGA**CUAGACAGCCAAC**UCA** | 12 | (((.....))).....  | N19(6r) |
| | **GGCACGA**CUAGACAGCCAAC**UCA** | 2 | (((.....))).....  | N23(4r) |
| | **GGCACGA**CUAGACAGCCAAC**UCA** | 1 | (((.....))).....  | N30(6r) |
| | **GGCACGA**CUAGACAGCCAAC**UCA** | 11 | (((.....))).....  | N37(7r) |
| | **GGCACGA**CUAGACAGCCAAC**UCA** | 1 | (((.....))).....  | N37(6r) |
| V | **GGCACGA**CAGCCCCGCC**UCA** | 2 | (((.....)))..... | N30(6r) |
| | **GGCACGA**CAGCUACCCGCCC**UCA** | 1 | (((.....)))..... | N30(6r) |
| | **GGCACGA**CAGCCACCGCCCC**UCA** | 1 | (((.....)))..... | N30(6r) |
| | **GGCACGA**CAGCCCCGCC**UCA** | 1 | (((.....)))..... | N37(7r) |

Categories I to V all have one stem loop and bear slightly different 3'-end tails. At first sight, these stem loops seem compatible with the existence of a catalytic site where ATP/GTP and an amino acid could bind, while the unstructured 3'-ends may possibly bind nearby. We established the distribution of the length of the single stranded 3' end in categories I - V (**Figure 3-8**). Interestingly, even though these sequences were selected independently from different libraries, the length of their unstructured 3'-end tails is not very disparate: most of them are between 3 and 10 nt and the distribution shows a narrow profile centered around 6 nt (**Figure 3-8**, left). We also analyzed the length distribution of the 3' end tail within the IV category, which is the most abundant (**Figure 3-8**, right). There are two peaks at 4 nt and 6 nt, while no representative is found with a 3' tail of 5 nt. This phenomenon seemed reminiscent of a structural effect observed with the Yarus 29-nt ribozyme (Lehmann et al. 2007) – one nucleotide change could cause a significant drop in the ribozyme activity and thus affect its enrichment among the population.

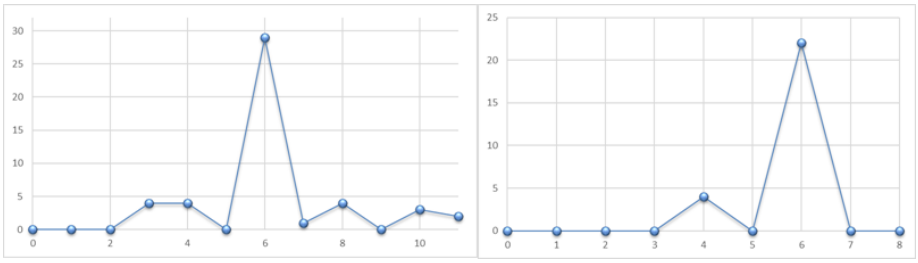



**Figure 3-8**. Left: Length distribution of unstructured 3'-end among category I to V. Right: Length distribution of unstructured 3'-end from category IV. X axis indicates the length. Y axis is the counting numbers.

Categories VI and VII are presented in the table below:

**Table 3-10**. Category VI and VII

| Cat. | Sequence | Copy | Secondary structure | Library |
|---|---|---|---|---|
| VI | **GGCACGA**GACACCUACCGCC**UCA** | 1 | (((...((....))...)))... | N30(6r) |
| | **GGCACGA**GACACCUACCGCC**UCA** | 7 | (((...((....))...)))... | N37(7r) |
| | **GGCACGA**GACAUCUAUCGCC**UCA** | 1 | (((...((....))...)))... | N37(7r) |
| VII | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 1 | ((.(((.....))).))....(((.....))). | N30(6r) |
| | **GGCACGA**CCAGCGUUCCUCUAGACCCCUACGUC**UCA** | 4 | ((.(((.....))).))...((((.....)))).. | N30(6r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUCU**UCA** | 5 | ((.(((.....))).))...((((.....)))).. | N30(6r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 12 | ((.(((.....))).))....(((.....))). | N37(7r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUC**UCA** | 6 | ((.(((.....))).))...((((.....)))).. | N37(7r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUCU**UCA** | 7 | ((.(((.....))).))...((((.....)))).. | N37(7r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACG**UCA** | 2 | ((.(((.....))).))....(((.....))). | N37(6r) |
| | **GGCACGA**CCAGCGUUCCUCUGGACCCCUACGUC**UCA** | 1 | ((.(((.....))).))...((((.....)))).. | N37(6r) |
| | **GGCACGA**CCAGCGUUCCCCUGGACCCCUACGUCU**UCA** | 1 | ((.(((.....))).))...((((.....)))).. | N37(6r) |

The above results suggested to us that our selection strategy to isolate self-aminoacylating ribozymes was successful. An additional investigation during which oligonucleotides involved in our selection protocol were compared to the selected sequences however revealed an unexpected phenomenon: the most frequently selected 23-nt RNA has a sequence identical to a portion of the 3' adaptor-primer used in the RNA ligation step (**Figure 3-9**).

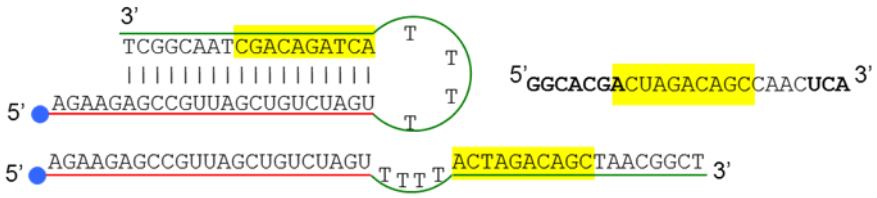



**Figure 3-9**. Sequence similarity between the adapter and the most abundant RNA target. The RNA sequences includes a part of the RT primer in a forward direction. The same sequence is highlighted. Blue circle: monophosphate; Red line: RNA sequence; Green line: DNA sequence.

This finding was a major disappointment for us. It showed that an unexpected phenomenon occurred during the selection, during which a "selfish mini-monster" (see **Section 1**) diverted our selection away from the targeted ribozymes. So far, we could not figure out what happened: it is absolutely not clear to us by which mechanism the identified sequence (**Figure 3-9**) could invade the libraries of our selection.

## 3.3 Tests of self-aminoacylation with the selected libraries

Before we knew about the "mini-monster" phenomenon, we decided to verify the aminoacylation activity of the pools of RNA and also test some specific sequences. We first tested library N19. Two parallel experiments were setup, in which the amino acids mix was present only in one of them (**Figure 3-10**). It was expected that the one with the amino acids should show a positive selection with the **OD** treatment while the experiment without amino acid should provide a negative signal in the absence of aminoacylation. **Figure 3-10** shows that both the **OD** and **DO** treatments gave a negative signal regardless of the presence of amino acids.

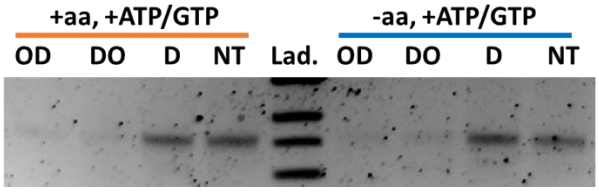



**Figure 3-10**. Test for aminoacylation with N19 library after 6 rounds selection.

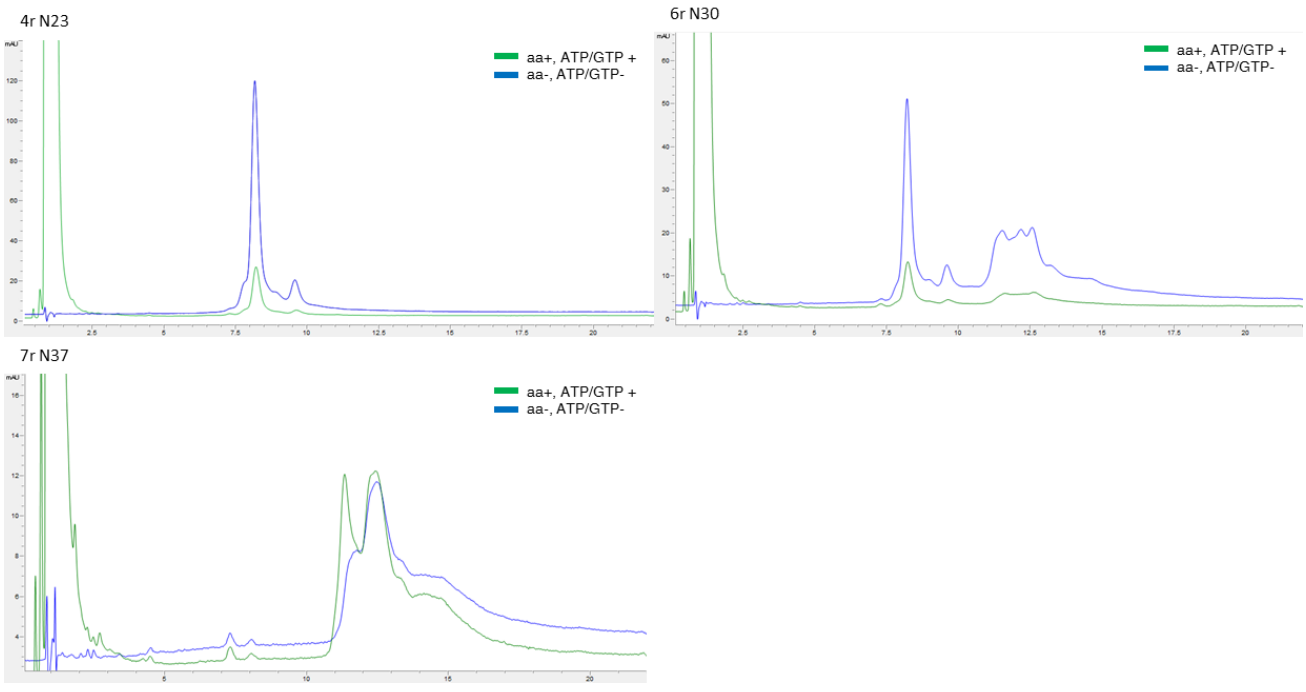



**Figure 3-11**. HPLC traces of libraries N23, N30 and N37 after incubation.

Since aminoacylated RNA are expected to have a different retention in the column of a liquid chromatography experiment (e.g. Lehmann et al. 2007), a time delay on reverse phase

HPLC is expected. We decided to use this different approach to further test the activity of the pools. Practically, pools of RNA can be injected on the HPLC right after an incubation reaction. We tested the library N23 after 4 cycles selection, N30 after 6 cycles selection and N37 after 7 cycles selection. No delay in the signals was apparent on the overlay of the two "+aa" experiment and "-aa" control (**Figure 3-11**).

An interesting result was however found when we tested the N37 library: while using a similar procedure as with the N19 pool (**Figure 3-10**), the signals were consistent with the protection of the 3'-end of some RNA due to the presence of the amino acids, suggesting that aminoacylation did occur (**Figure 3-12**). This result still needs to be reproduced.

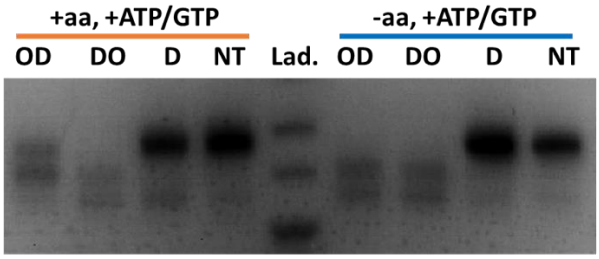



**Figure 3-12**. Verification of aminoacylation for the library N37 after 7 rounds selection.

Besides, we also choose a few specific sequences to test them individually with HPLC. In total 23 specific sequences were divided into 5 batches for the tests (**Supporting information**). No clear signal could be detected by HPLC. **Figure 3-13** shows two examples of the HPLC analysis.

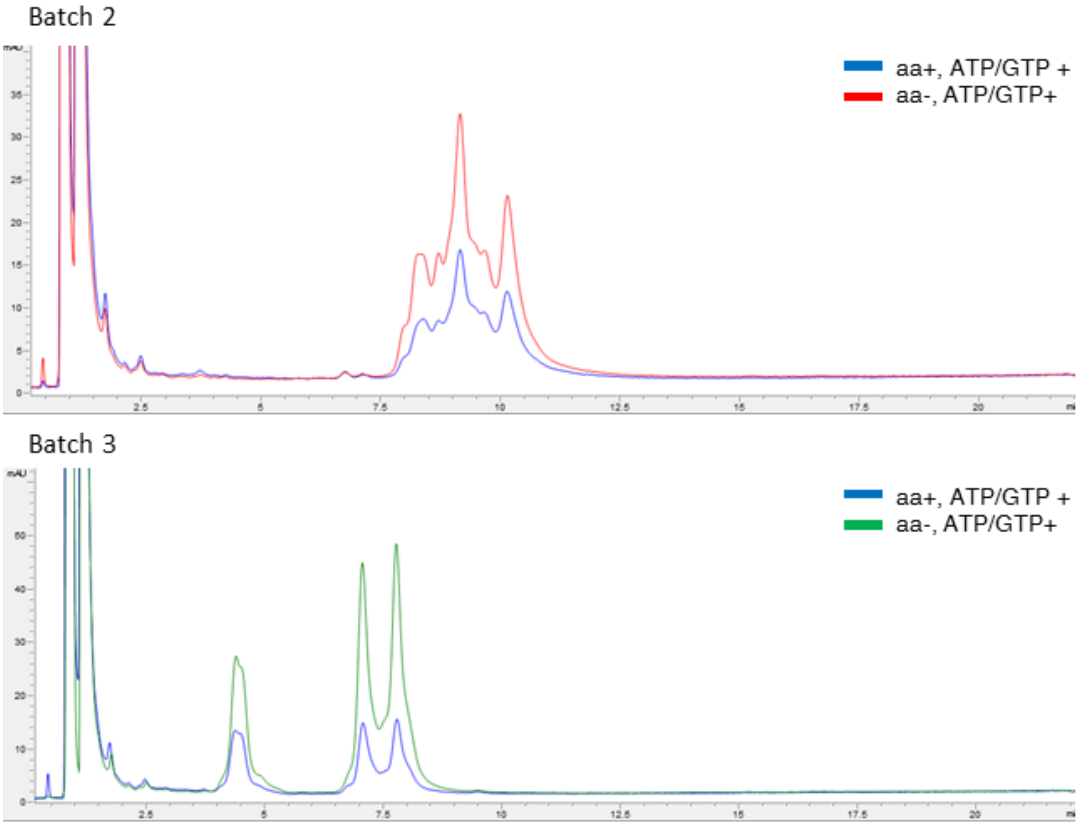



**Figure 3-13**. HPLC analysis for specific sequence aminoacylation test.

## 3.4 Conclusion

We established a SELEX protocol to isolate self-aminoacylating ribozymes while using optimized OD-DO assays. A carefully designed hybrid RNA adapter was used to achieve efficient RNA ligation. Moreover, this hybrid adapter can significantly prevent unspecific priming during the reverse transcription step. In this protocol, we also used primers with small overhangs to minimize RNA size heterogeneity during the selection, and to avoid an unmanageable step (the ligation of a forward adaptor). Combining these techniques, we could limit the size of the conserved regions down to 10 nt.

The available substrates (a mixture of 19 amino acids with ATP/GTP cofactors) required the targeted ribozymes to activate the amino acids with either ATP or GTP and realize transesterification to the 3'-end. Although we could finalize a working SELEX procedure while minimizing the constant tracks at both the 5' and 3' ends, tests with the libraries obtained after 6 to 7 cycles failed to reveal any clear activity. Furthermore, we identified an unexpected phenomenon that prevented our selection protocol to isolate the targeted ribozymes. Besides these considerations, there are several possible reasons that may have cause the failure of our approach:

1) The cocktail buffer is inappropriate for aminoacylation. It is the risk of a "blind selection" existing in all *in vitro* SELEX experiment. Further investigation could allow us to improve the composition of the buffer.
2) Periodate oxidation could not eliminate hundred percent of non-aminoacylated RNA. Some of the RNA molecules, for instance the ones with strong secondary structures could possibly resist oxidation, will surely be selected.
3) Our idea was to select a class of small ribozymes. Our initial pool may not be large enough to fulfill the requirements of the reactions being investigated. Ribozyme being able to catalyze both amino acid activation and transesterification may require larger structural motifs. In order to promote nucleophilic attack, RNA molecules should be able to tightly bind amino acid and ATP substrates. and bring them to a reaction center, meanwhile the 3'-end of RNA should also be able to reach this center to accept the activated amino acid. This sophisticated binding system may require a longer RNA

chain and larger initial pool of RNA to provide a higher diversity to the selection.

4) Given the unexpected phenomenon that occurred during selection, the SELEX protocol should be redesigned to prevent the possibility of these effects.
5) At last, some additional experiments should still be undertaken to verify whether some selected sequences could still promote aminoacylation.

# Chapter II Identification of aminoacylated RNA from *E. coli*

## 1. Introduction

As part of our work on aminoacylated RNA, we wanted to develop a robust **OD-DO** procedure to not only investigate the possibility of new self-aminoacylating ribozymes (**Chapter I**), but also to analyze total RNA from different species, and find out whether aminoacylated RNA other than tRNA and tmRNA could be identified in living cells. For this purpose, we choose to work with total RNA from *Escherichia coli*, a well-studied Gram-negative bacterium. Our approach relies on total RNA extraction under non-deacylation conditions (essentially while keeping a low pH during the procedures). After extraction, RNA samples follow different series of treatments: **OD** (oxidation-deacylation), **DO** (deacylation-oxidation), **D** (deacylation only) and **NT** (no treatment) as is described in last chapter. Treated samples are processed to generate cDNA banks through the addition of adaptors, and these banks are sent to a company for deep sequencing using the Illumina pair-end sequencing technology. A comparison between batches is expected to provide the information to identify aminoacylated species.

In contemporary cells, tRNA is the prototypical molecule of aminoacylated RNA. Its primary function is to participate in translation on the ribosome, where proteins are being built according to the rules of the genetic code. More than 20 different variants of this molecule are specifically aminoacylated by aminoacyl-tRNA Synthetase (Giegé and Springer 2012). In bacteria, aminoacylated tRNA not only participate in translation; they are also involved in cell wall metabolism (Raina and Ibba 2014). Furthermore, prokaryotic cell uses several strategies to maintain appropriate levels of aminoacylation through the use of different sensors such as the T-box riboswitch (mainly in Firmicutes). The study of the level of aminoacylation with an **OD-DO** methodology could thus highlight new information about this essential actor in cell metabolism. Furthermore, because tRNA molecules constitute a large fraction (15%) of total RNA in bacteria (Uzman 2001), tRNA is an obvious candidate to validate and calibrate our protocol.

There are several indications that other types of RNA molecules could be aminoacylated in cells. In plants, it has been shown that some RNA viruses bear transfer RNA-like structures

(TLSs) that require aminoacylation to achieve a full cycle of infectivity (Dreher 2009). Another interesting fact is the aminoacylation property of tRNA minihelixes (Francklyn and Schimmel 1990): in vitro experiments showed that small stem-loop mimicking the acceptor stem of tRNA can be aminoacylated by the aminoacyl-tRNA synthetase, although at low efficiency. It is not known whether this type of aminoacylation could also spontaneously occur *in vivo*, and have functional purposes.

By using the deep sequencing technology, our investigation could highlight new RNA candidate for aminoacylation, and further explore the rules that govern the aminoacylation levels of tRNAs.

## 2. Method

Our approach relies on the careful preparation of total RNA extracted from *E. coli* followed by oxidation and control protocols: **OD, DO, D** and **NT** (Supporting information ***OD-DO treatment***). Similar to the SELEX experiment, **OD** treated sample are expected to provide the reads of aminoacylated RNA. **DO** processed samples is a negative control that should not provide any substantial specific signal except for RNA with 3' methylation (one unknown phenomenon in bacteria) (Motorin and Helm 2011) or if **DO** treatment may still leave non-oxidized 3' terminations and degraded small RNA fragments during the bench manipulation. **D** processed samples should make any RNA available for sequencing (including aminoacylated RNA) while **NT** samples constitutes the untreated positive control.

### 2.1 Primary attempt with a standard Illumina RNA-seq protocol

In our first attempt to acquire deep-sequencing data with treated samples of total RNA, a standard RNA-seq bank preparation protocol was followed (TruSeq Small RNA kit, Illumina). We provided treated RNA samples to the Plateforme de Séquençage facility at Gif-sur-Yvette (Imagif). A home-made pre-purification protocol by gel electrophoresis, established in collaboration with the technician at the Facility (Erwin Vandijk), was implemented to obtain three samples belonging to different size categories: 10-50 nucleotides, 50-100 nucleotides and 100-250 nucleotides. It is in order to normalize the weight of each category, and prevent the 50-100 nt category (containing most of the tRNA) to become overwhelmingly dominant. For each category, an RNA spike of appropriate size was designed, and a normalized mixture of the three spikes was added to the processed samples before gel purification. Gel purification of the **OD, DO, D** and **NT** samples resulted in 12 different samples, each of them being labelled with an index sequence corresponding to the Illumina sequencing requirement during bank preparation.

Bank preparation was achieved by the consecutive ligation of universal adaptors (with T4 RNA ligase) to both the 5'-end and the 3'-end (**Figure 2.1**). A polynucleotide kinase (PNK) treatment was performed to ensure that the 5'-end of the RNA samples had a monophosphate.

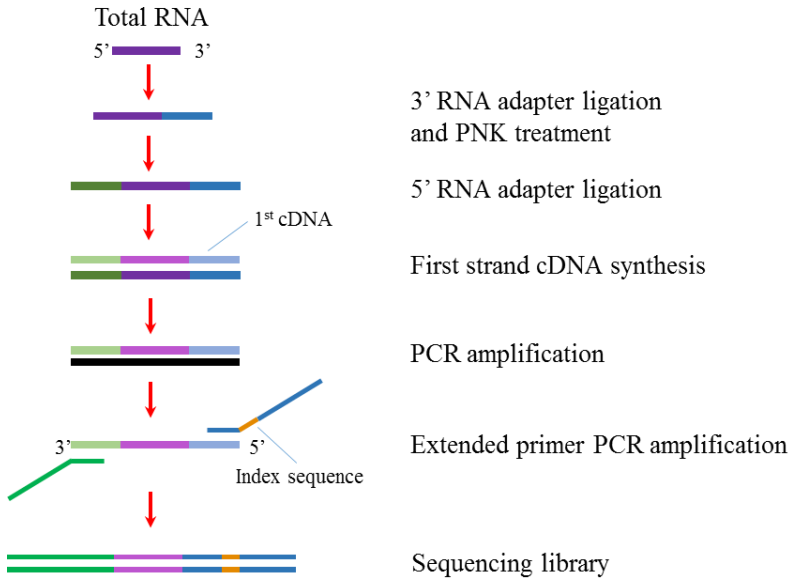



**Figure 2.1**: Illumina TruSeq Small RNA library standard preparation (work flow). Dark blue: 3' RNA adapter and lighter color represents the complementary DNA sequence; Dark green: 5' RNA adapter and lighter color indicates the complementary DNA sequence; Purple: target RNA fragments.

### 2.1.1 First sequencing Results

#### Sample preparation

Our former total RNA samples were prepared by our collaborators from ETH in Zurich (T. Grentzinger and A. Marchais). They provided us with total RNA samples from *E. coli* BL21 strain extracted during exponential phase with Trizol, in a final form of frozen aliquots of 5 $\mu$l at a concentration of 3 $\mu$g/$\mu$l. They were shipped in dry ice, and immediately stored at -68°C upon receiving. An analysis with bioanalyzer however revealed that the quality of these samples was not high enough for deep sequencing. We decided to prepare new samples by ourselves from the *E. coli* BL21 strain kindly provided by the laboratory of Philippe Bouloc (I2BC).

Total RNA from *E. coli* BL21 was extracted at $OD_{600}$ point: 0.24, 0.47, 1.36 as well as an overnight culture, the final sample was constituted by an equal amount of each extraction. This sample was split into four aliquots, each of them (~4 $\mu$g total RNA) being processed through either the **OD, DO, D** or **NT** protocol (Supporting information, total RNA extraction). Then, these samples were treated with DNase I, re-suspended in 15 $\mu$l DEPC-water and quantified (**Table 2.1**). An equal amount of spike was added in each sample (see below) and the quality of the RNA was checked with Bioanalyzer (**Figure 2.2**). The presence of ribosomal RNA

confirmed that no excessive degradation occurred during the treatments, and the samples were sent for bank preparation at the *Plateforme de Séquençage* (Imagif, Gif-sur-Yvette).

**Table 2.1**. Concentrations of total RNA recovered from **OD, DO, D** and **NT** treatments

| | **OD** | **DO** | **D** | **NT** |
|---|---|---|---|---|
| *c* (ng/$\mu$l) | 399.2 | 356.1 | 415.4 | 409.5 |

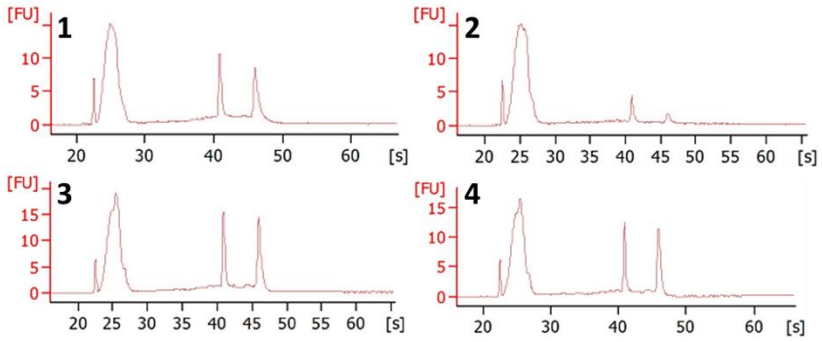



**Figure 2.2**. Bioanalyzer traces of treated RNA sample: 1-4 number represents **OD, DO, D** and **NT** respectively. A diluted sample from the recovered samples were injected for analysis.

**Spikes added to the samples:**

Spikes are exogenous RNAs that allow the normalization of signals obtained from a series of samples. Three spikes of a size comprised within the defined categories (10-50 nt, 50-100 nt and 100-250 nt) were added in equal amount to the four treated samples (**Table 2.2**). RNAspike_25 and RNAspike_65 are chemically synthesized RNA (IDT), and correspond to initial portions of RNAspike_131. RNAspike_131 is the transcript of the control template of the MEGAshortscript™ kit (Ambion), a fragment of human 18S rRNA. This spike was *in vitro* synthesized and HPLC purified.

**Table 2.2**. Spikes used for normalization

| | | |
|---|---|---|
| RNAspike_25 (for 10-50 nt category) | Size: 25 nt | MW: 8172.9 g/mol |
| GGG AGA GAG GGC UGC UGU UCU AGA G | | |
| RNAspike_65 (for 50-100 nt category) | Size: 65 nt | MW: 20991.6 g/mol |
| GGG AGA GAG AGA GAA UUA CCC UCU ACG CUA UUG GAG CUG GAA UUU CCG CGG CUG CUG UUC UAG AG | | |
| RNAspike_25 (for 10-50 nt category) | Size: 25 nt | MW: 8172.9 g/mol |
| GGG AGA GAG AGA GAA UUA CCC UCA CUA AAG GGA GGA GAA GCU UAU CCC AAGAUC CAA CUA CGA GCU UUU UAA CUG CAG CAA CUU UAA UAU ACG CUA UUG GAG CUG GAA UUU CCG CGG CUG CUG UUC UAG AG | | |

A Stock solution of the three spikes mixture at approximately equal molar concentration (1:0.94:0.85) was prepared, and the mass concentration of the solution was established by *nano*-drop (2269 ng/$\mu$l). This stock was diluted to 1/200 and 2 $\mu$l were added to each sample, corresponding to a ratio of approximately 5% of spike per RNA sample (in mass).

**Bank preparation**

The samples were gel purified at the *Sequencing Platform Facility* at Gif-sur-Yvette, which allowed a splitting into the three defined size categories (10-50, 50-100 and 100-250 nt). Importantly, no RNA fragmentation was applied, a procedure that is often part of an RNA-seq bank preparation. The 12 samples resulting from this procedure were individually labelled during bank preparation by a specific index sequence on the 3' Universal Adaptor of the TruSeq Small RNA kit (Illumina) (**Table 2.3**). The 12 cDNA banks were processed by Beckman-Coulter (USA) for deep sequencing using the Illumina HiSeq2500 pair-end 2x125 technology.

**Table 2.3**: cDNA banks prepared by the *Sequencing Platform Facility* at Gif-sur-Yvette.

| Bank name | Inserts size (nt) | Index | Concentration (ng/µl) | Volume ($\mu$l) |
|---|---|---|---|---|
| OD-A | 100-250 | 16 | 13,6 | 18 |
| OD-B | 50-100 | 17 | 6,37 | 18 |
| OD-C | 10-50. | 18 | 27 | 18 |
| DO -A | 100-250 | 10 | 0,74 | 18 |
| DO -B | 50-100 | 11 | 4,11 | 18 |
| DO-C | 10-50. | 12 | 21,9 | 18 |
| D-A | 100-250 | 19 | 1,53 | 18 |
| D-B | 50-100 | 20 | 7,78 | 18 |
| D-C | 10-50. | 21 | 27,9 | 18 |
| NT-A | 100-250 | 13 | 3,36 | 18 |
| NT-B | 50-100 | 14 | 8,49 | 18 |
| NT-C | 10-50. | 15 | 17,4 | 18 |

**Deep-sequencing results**

On average, 20.75 million reads per bank (min = 13 Mio; max = 25 Mio) were sequenced, each transcript being constituted by R1 and R2 reads. An analysis of these transcripts (FastQC report) revealed an excellent overall quality.

Demultiplexing, adapter removal and mapping was achieved with *Bowtie* (Langmead et al. 2009). R1 and R2 reads were mapped on a reference genome of *E. coli* BL21. During this

operation, we realized that a large fraction of the reads could not map on *E. coli* BL21 genome, and did neither match our spike sequences. A BLAST search revealed that our samples became contaminated with some material from *Saccharomyces cerevisiae* (chromosome XII) (**Table 2.4**). An investigation could not reveal the origin of this contamination. A sufficient number of *E. coli* reads were however available for mapping, and allowed us to perform a preliminary analysis, and compare the reads found in some selected tRNA between the four tested conditions.

**Table 2.4**. Mapping of sequencing reads onto *E. coli* and *S. cerevisiae* genomes

| **Bank name** | Total number of reads | % mapping against *E. coli* BL21 genome | % mapping against *S. cerevisiae* genome |
|---|---|---|---|
| OD-A | 18001500 | 87.8 | 11.1 |
| OD-B | 24566899 | 11.2 | 83.1 |
| OD-C | 23488573 | 24.5 | 79.7 |
| DO-A | 12521884 | 79.3 | 17.3 |
| DO-B | 19888979 | 14.2 | 73.3 |
| DO-C | 21609403 | 26.7 | 80.1 |
| D-A | 16934962 | 93.5 | 5.8 |
| D-B | 20650043 | 13.6 | 80.3 |
| D-C | 23694044 | 23.9 | 78.4 |
| NT-A | 22409868 | 96.2 | 3.2 |
| NT-B | 21615843 | 17.4 | 76.9 |
| NT-C | 20763821 | 25.1 | 78.1 |

We first established the number of reads mapping on the spikes in order to normalize the signals observed in each condition. Although the number of reads in the 50-100 nt category was reduced due to *S. cerevisiae* contamination, most tRNA sequences would fall into this category. We therefore decided to examine it in priority.

**Comparison between OD, DO, D and NT signals in the 50-100 nt category**

An examination of the files with the *IGV* software (Robinson et al. 2011; Thorvaldsdottir et al. 2013) could immediately reveal an unexpected result: it appears that the oxidation treatments did not reduce the reads coverage observed in **OD** and **DO** samples, compared with the **D** and **NT** controls. **Figure 2.3** illustrates the results with a window analysis encompassing a methionine tRNA operon (*MetZ, MetW* and *MetV*). While taking into account the spike correcting factor, all three tRNAs are covered by approximately the same number of reads regardless of the **OD, DO, D** or **NT** treatment (**Table 2.5**).

**Table 2.5**. Spike normalization

| | | RNAspike_65 | | | |
|---|---|---|---|---|---|
| Treatment | **MetZ-MetW-MetV** maximum coverage | Maximum coverage (mc) | Sample conc. (sc) | Normalization factor ($mc_{min}$/mc)*(sc/$sc_{min}$) | **MetZ-MetW-MetV** normalized coverage |
| **OD** | 10726 | 78246 | 399.2 | 0.5655 | 6066,04 |
| **DO** | 10729 | 59684 | 356.1 | 0.6614 | 7095,98 |
| **D** | 6589 | 39474 | 415.4 | 1.1665 | 7686,24 |
| **NT** | 7066 | 45216 | 409.5 | 1.0039 | 7093,73 |

The situation observed with $tRNA^{Met}$ is not an exception: although there is a variation of the relative proportions of read counts between treatments, an inspection of the read coverages for other tRNA shows that this phenomenon is general.

**Figure 2.3** shows why **OD** and **DO** treatments fail to remove tRNA reads from the corresponding banks: although RNA fragmentation was purposely not implemented after the treatments, the large majority of the reads (regardless of the RNA identity) are fragments of full-length RNA molecules, revealing that fragmentation still significantly occurred during bank preparation, i.e. after oxidation and deacylation treatments. Fragmentation may generate free 3'-ends available for adaptor ligation; these fragments therefore may be included in the banks.

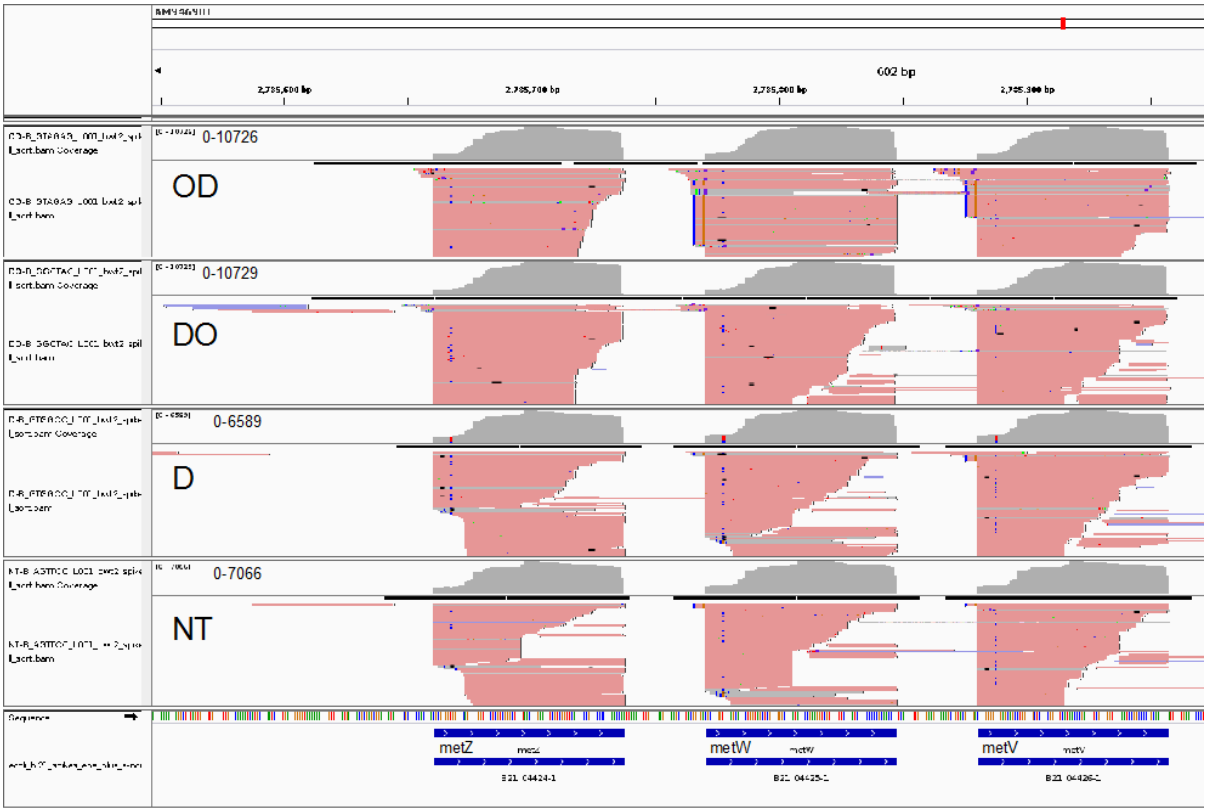



**Figure 2.3**. **OD, DO, D** and **NT** reads mapping a methionine tRNA operon (*MetZ, MetW* and *MetV* genes, in dark blue). Reads (in brick color; alternate colors highlight read mismatches) are classified with the longest ones on top. Only the upper reads (constituting a small fraction of all reads) are visible. Gene coverage in shown on top (in grey).

A striking feature, likely related to this undesired phenomenon, is that full-length reads are almost absent: they only constitute a few percent of the total number of reads for a given molecule (**Figure 2.4**). Why are there so few entire tRNA molecules in the banks? In addition to fragmentation, a plausible explanation is that modified bases present on full-length tRNA prevent the progression of the reverse transcriptase during cDNA synthesis (Motorin et al. 2007). This final step of bank preparation (**Figure 2.1**) has to be fulfilled in order for a considered read to undergo final PCR.

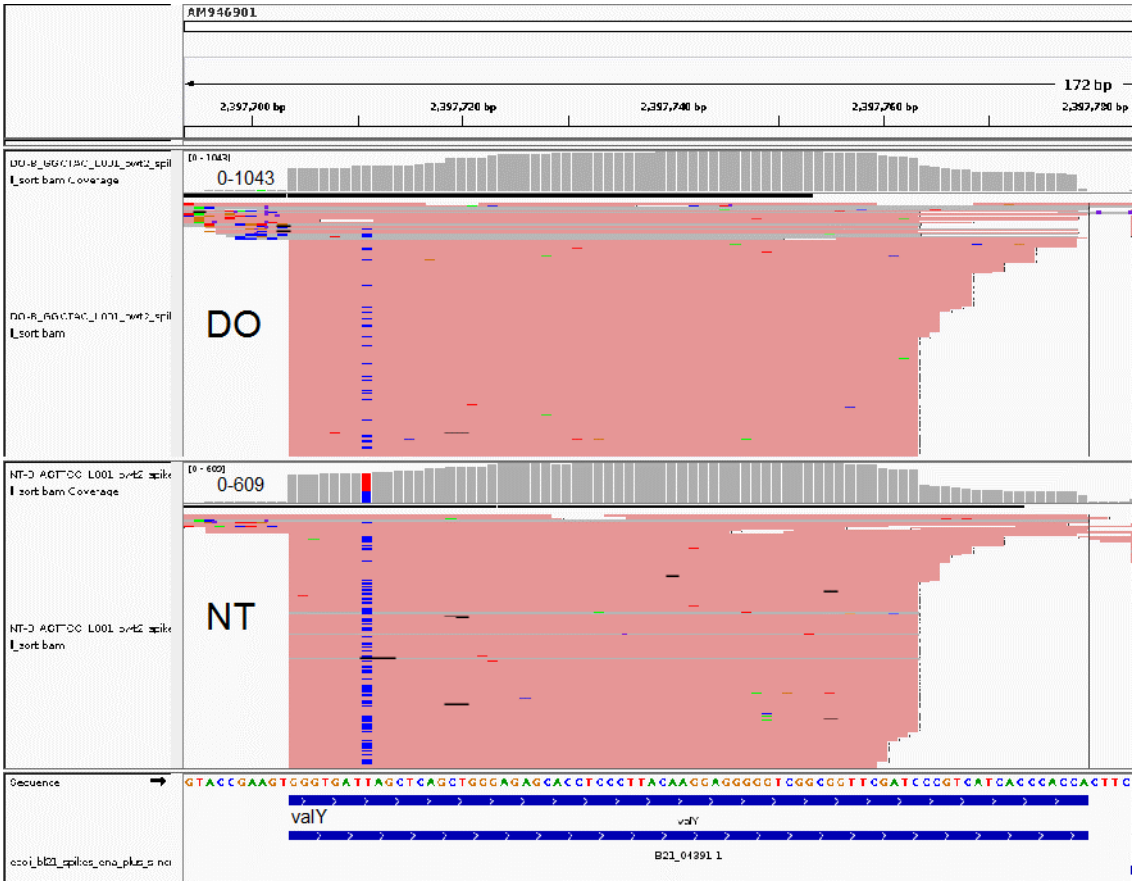



**Figure 2.4**. **DO** and **NT** reads mapping the *valY* gene (a valine tRNA, in dark blue). Reads (in brick color; alternate colors highlight read mismatches) are classified with the longest ones on top. Only the upper reads (constituting a small fraction of all reads) are visible in the two **DO** and **NT** windows. Spike normalization (**DO** = 1043 x 0.6614 = **690**; **NT** = 609 x 1.0039 = **611**. See **Table 2.5**) reveals that both treatments have a similar coverage. Gene coverage in shown on top (in grey).

An undesired consequence of this phenomenon is the following: because only fully modified tRNA are (essentially) aminoacylated by the aminoacyl tRNA synthetase, our bank preparation protocol could not quantitatively access the information about the level of aminoacylation of tRNAs.

We still wanted to know whether our present protocol could provide some information about tRNA aminoacylation. An inspection of the 3'-end of reads mapping tRNA genes revealed an unexpected signature of the **DO** reads when compared with those undergoing other treatments: instead of a regular –CCA 3' termination, the large majority of these reads lack the terminal adenosine (–CC 3' termination). **Figure 2.5** shows in fact that with **OD, D** and **NT** treatments, already half of the reads lack the terminal adenosine. Although we could not explain this feature, it appears that the nature of the termination might be connected with aminoacylation, and may still allow us to quantitatively compare **OD** and **DO** samples.

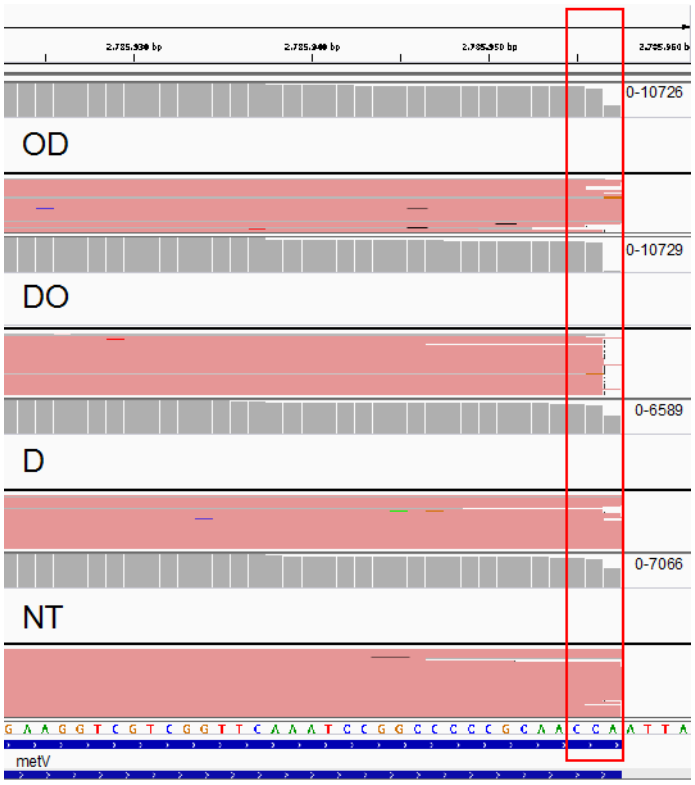



**Figure 2.5**. **OD, DO, D** and **NT** reads mapping a *MetV* gene (methionine tRNA). This magnification of **Figure 2.3** reveals that almost all CCA 3' end of **DO** treated tRNA reads lack the terminal adenosine, while this base is missing in approximately half of the reads with the other treatments.

While only considering the reads finishing with either –CCA 3' or –CC 3', we established the fraction of reads ending with –CCA for all tRNA genes (**Figure 2.6**). The analysis shows that this ratio has a high dispersion for the **NT** control (**Figure 2.6**, A). Because we don't know whether an A76 removal reflects the situation with full-length tRNA or tRNA fragments (see above), any interpretation would be hazardous. While this effect needs to be further investigated, it can be mentioned that A76 removal by RNase Z has been reported in *E. coli* (Ezraty et al. 2005; Takaku and Nashimoto 2008), and could play a role in translation regulation.

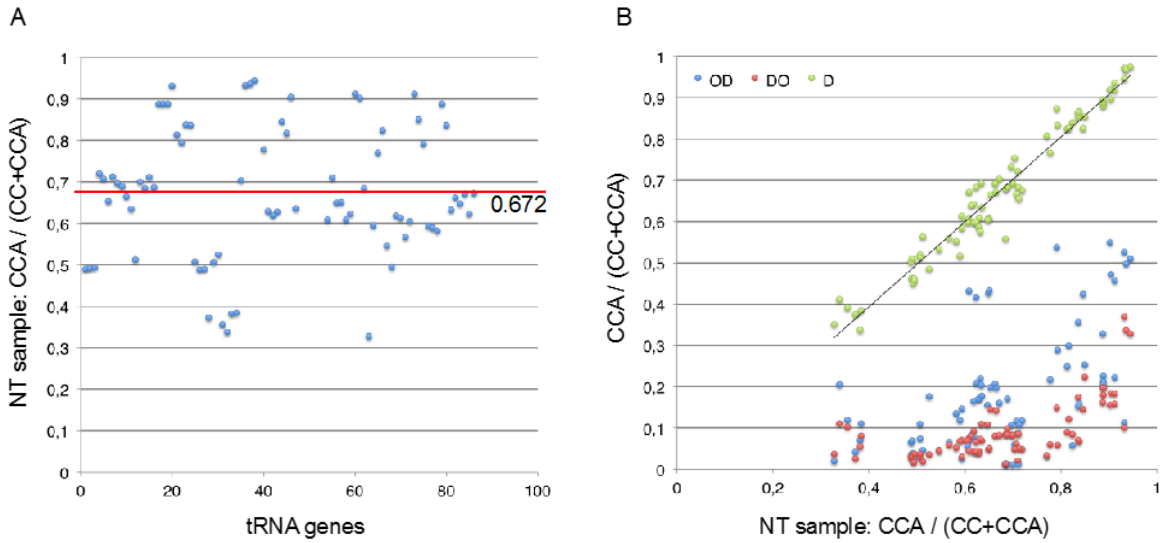



**Figure 2.6**. Nature of the 3' termination in *E. coli* tRNA reads. A) Dispersion of the CCA/(CC+CCA) ratio in all tRNA genes, **NT** sample (in alphabetical order): [alaT, alaU, alaV, alaW, alaX, argQ, argU, argV, argW, argX, argY, argZ, asnT, asnU, asnV, asnW, aspT, aspU, aspV, cysT, glnU, glnV, glnW, glnX, gltT, gltU, gltV, gltW, glyT, glyU, glyV, glyW, glyX, glyY, hisR, ileT, ileU, ileV, ileX, leuP, leuQ, leuT, leuU, leuV, leuW, leuX, leuZ, lysQ, lysT, lysV, lysW, lysY, lysZ, metT, metU, metV, metW, metY, metZ, pheU, pheV, proK, proL, proM, selC, serT, serU, serV, serW, serX, thrT, thrU, thrV, thrW, trpT, tyrT, tyrU, tyrV, valT, valU, valV, valW, valX, valY, valZ]. Due to low number of read counts (< 40), [IleX, lysQ, lysT, lysV, lysW, lysY, lysZ] values are not shown. The average value (red line) is 0.672. B) Correlation of the CCA/(CC+CCA) ratio between untreated sample (**NT**) and the **OD**, **DO** and **D** samples. The linear regression of the **D** sample has a slope of ~1.0.

The high correlation observed between **D** and **NT** samples (**Figure 2.6**, B) demonstrates that the deacylation treatment (pH 9.0) does not affect the integrity of the 3'-ends of the reads, but also that tRNA from the **NT** were still exposed to deacylation conditions, which make their 3'-ends available for ligation. Considering the **OD** and **DO** cases, the dispersion observed in **Figure 2.6** (B) is consistent with the hypothesis that the presence of an amino acid prevents the removal of the terminal A76 nucleotide, that may occur after oxidation at high pH. It suggests that during bank preparation, our samples experienced conditions in which the pH was sufficiently high to achieve $\beta$-elimination before adaptor ligation. $\beta$-elimination is a common method to remove one nucleotide from the 3'-end of RNA by means of periodate oxidation and subsequent deacylation (**Chapter I**, OD-DO assay). It is however not clear to us why this effect appears in **D** and **NT** samples and seems to only affect A76 but not C75 (half of the reads in **D** and **NT** samples end with C75).

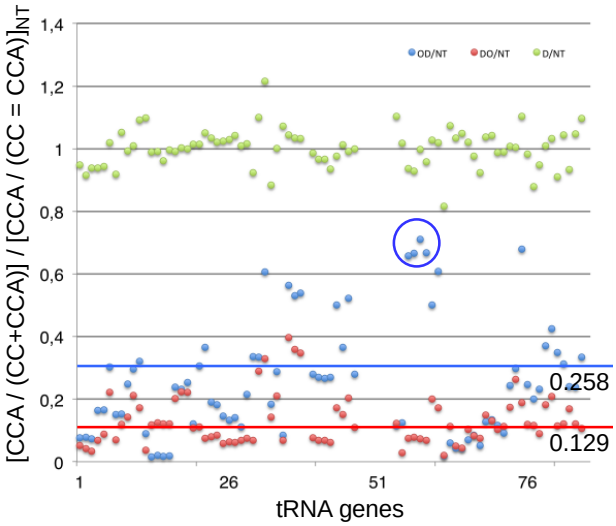



**Figure 2.7**. CCA/(CC+CCA) ratios for **OD, DO** and **D** samples normalized by the ratios of the **NT** condition in all tRNA genes (**Figure 2.6**). This plots further highlights the neutral effect of the **D** treatment (the average value of the ratio is ~1). The four circles dots correspond to the **OD** values of methionine tRNA genes *metV*, *metW*, *metY* and *metZ*.

The values observed in **Figure 2.7** confirm that **DO** reads are those with the highest proportion of –CC 3' termination. A value close to zero is "expected" for these reads if oxidation would lead to a complete removal of the A76 residue. Interestingly, the average values for **OD** (0.258) and **DO** (0.129) suggest the possibility that among –CCA reads only, half of them (0.258/0.129 = 2) are aminoacylated on average. It can also be seen that a higher dispersion of the values is observed with **OD** reads compared to **DO**, consistent with the possibility of a high variability of the level of aminoacylation between specific tRNA. A high value of the **OD** ratio is observed for almost all methionine tRNA (blue circle), suggesting that the aminoacylation level for these tRNA is particularly high.

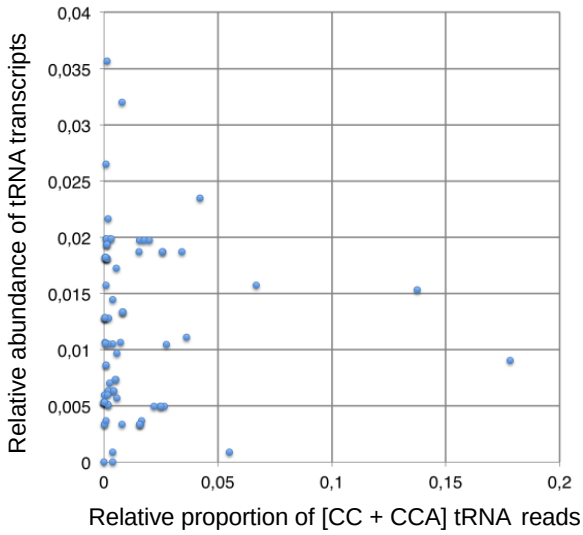



**Figure 2.8**. Correlation between the relative proportion of tRNA –CC and –CCA reads from our RNA-seq analysis (**NT** sample) and relative tRNA abundance determined *in vivo* [data from (Dong et al. 1996a)]. 85 tRNA genes are considered.

Finally, we wanted to see whether the proportions of the [–CC + –CCA] read counts between different tRNA genes could reflect the relative abundance of these tRNA observed in *E. coli*. The comparison of our data with the tRNA abundance experimentally determined by Kurland and coworkers using 2D high resolution gel analysis (Dong et al. 1996b) show that our data are far from being representative (**Figure 2.8**). An examination of the **NT** read counts for individual tRNA reveals that lysine and isoleucine tRNA reads are almost absent (<10 reads) when some leucine tRNA have more than 20,000 reads. Although the various effect(s) of base modification on reverse transcription efficiency is unknown, it is quite plausible that it may explain these discrepancies (Zheng et al. 2015).

In conclusion, a standard RNA-seq protocol is not appropriate for the study of the aminoacylation level of RNA molecules due to fragmentation during bank preparation. In addition, the presence of modified bases on tRNA may not allow a quantitative assessment of these species in solution. This second issue has recently been addressed (or at least been partly overcome) with a protocol during which methylations are removed prior to bank preparation (**Figure 2.9**) (Zheng et al. 2015). In this improved procedure, the use of a template-switching polymerase (TGIRT) also enables a higher amount of cDNA resulting from aborted reverse transcription events to still be included in the bank. This approach is however problematic in our study of aminoacylation because it requires demethylase treatments in alkaline buffers with

high amounts of $Mg^{2+}$ and $Fe^{2+}$ (Wilusz 2015; Zheng et al. 2015), during which deacylation will occur. In the following Section, we propose a new method of small RNA sequencing to specifically evaluate the aminoacylation level of any RNA transcript from total RNA.

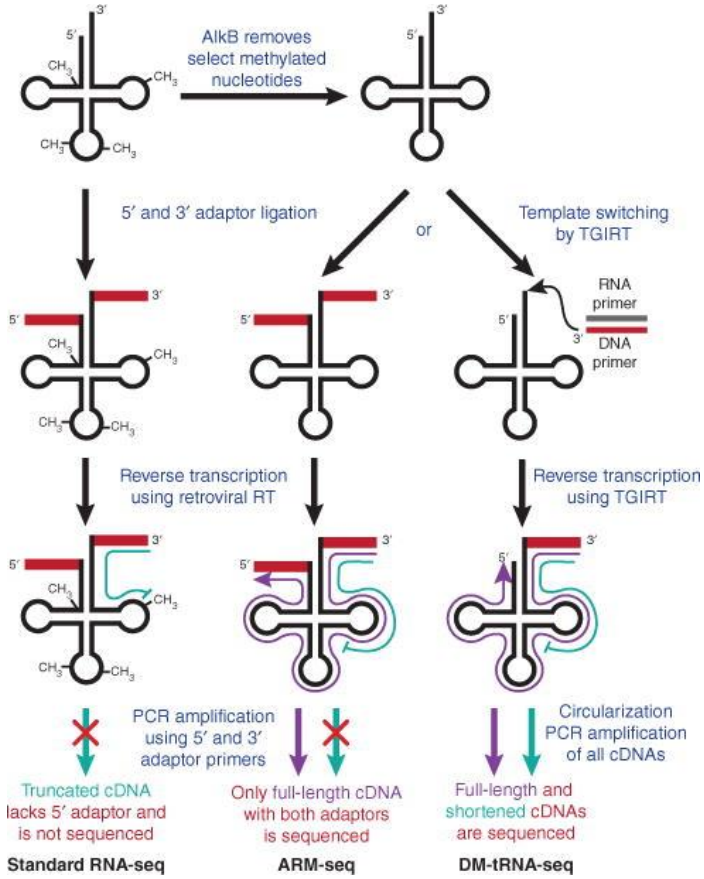



**Figure 2.9**: Removing methylated nucleotides to sequence tRNAs. Whereas reverse transcriptase (RT) cannot extend through a fully modified tRNA (left), treatment with AlkB removes select methylated nucleotides to allow efficient reverse transcription and deep sequencing (center and right). Full-length and truncated cDNAs are shown in purple and green, respectively (Zheng et al. 2015).

## 2.2 New RNAseq protocol for the study of the aminoacylation level of RNA transcripts with modified bases

The identification of issues in the study of aminoacylated RNA species with a standard RNA-seq protocol allows us to propose a new protocol optimized for that purpose. Important considerations are the following:

- In order to establish a reliable comparison between **OD, DO, D** and **NT** samples, our previous investigation showed that 3’ adaptor ligation should occur immediately after these treatments while gel purification and PNK treatment steps must be avoided.

- Since we are interested in the aminoacylation level of the 3’ end of these molecules, it is not necessary to obtain full-length cDNA of the RNA transcripts since 5’-end sequence information is not essential. A size of at least ~20 nt is however desirable in order to achieve proper mapping of these transcripts on the genome.

Since the reverse transcriptase has high probability to drop and generate truncated cDNA when it encounters modified bases (Zheng et al. 2015), a usual two-step adapter ligation cannot be performed. In order to add a 5’ adapter, this protocol combines reverse transcription with the RNA 5’RACE technique (Frohman et al. 1988; M.A. Innis, D.H. Gelfand 1990). **Figure 2.10** shows the work flow of the new protocol. Treated RNA are reverse transcribed into cDNA right after 3’ adapter ligation. A gel purification of the DNA/RNA duplex enables us to remove the excess of 3’ adapter and select transcripts within a size range of interest (from 35 to 300 nt). Full-length and truncated cDNA then undergo poly-A tailing with a **T**erminal **d**eoxynucleotidyl **T**ransferase (TdT) in conditions for which a tail of the order of 20 to 30 adenosine residues is added. This tail allows a reverse adapter comprising a tail of 17 dT at the 3’ end to prime the 3’ end of these cDNA. A two-step PCR is achieved with a reverse universal primer and a forward indexed primer, which finalizes bank preparation.

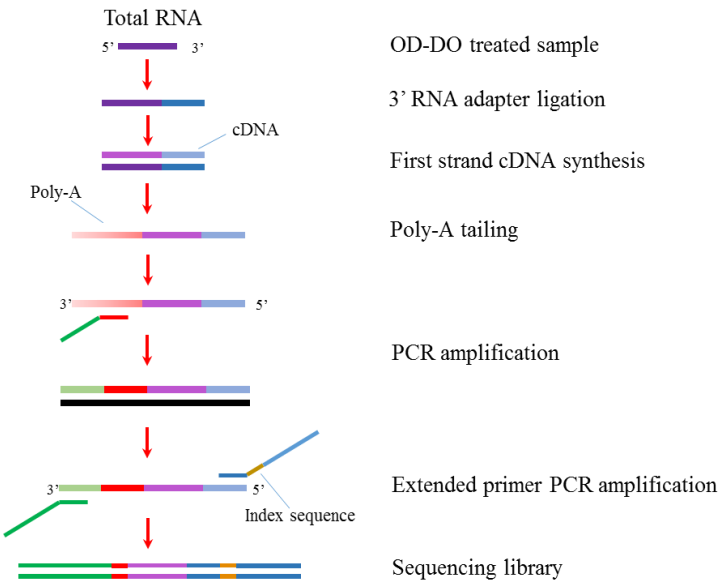



**Figure 2.10**. Work flow of homemade aminoacyl RNA sequencing library preparation. Dark purple: treated RNA samples; Dark blue: pre-adenylated DNA adapter; Red: poly-A tail; Green and light blue: Illumina TruSeq required oligos.

### 2.2.1 Experiment

#### Sample preparation

For this second round of deep-sequencing analysis, we decided to work with *E. coli* MG1655 instead of *E. coli* BL21 on suggestion of Philippe Bouloc (I2BC). *E. coli* MG1655 being a well-studied strand of *E. coli*, new results will be easier to relate to existing data. Furthermore, we decided to establish a biological triplicate of our total RNA extraction at a single $OD_{600}$ value of 0.4. An additional sample was established from an equal mixture of 4 extractions collected at $OD_{600}$: 0.1, 0.4, 1.0 and overnight culture. As before, all extractions are performed in cold acidic phenol to avoid deacylation. The 3+1 samples were split into four aliquots (~4 $\mu$g total RNA each) for the **OD, DO, D** and **NT** protocols (supporting information *OD-DO treatment*). All 16 samples were DNase treated, re-suspended in pure water (20 $\mu$l) and quantified. The established concentrations were used to determine the volume of a spike solution (RNAspike_65 only; **Table 2.2**) to be added in each sample in order to reach a final 2% of spike (in mass). The quality of the samples was verified by gel electrophoresis (**Figure 2.11**).

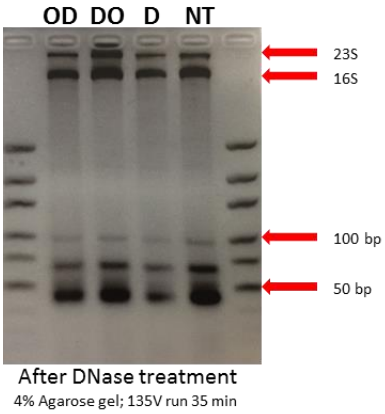



**Figure 2.11**. Verification of sample integrity after treatment by gel electrophoresis. After DNase treatment, around 300 ng total RNA was loaded on gel. The presence of clean 23S, 16S rRNA bands as well as small RNA part show that no significant RNA degradation occurred during treatment.

#### Bank preparation

Following the new workflow shown in **Figure 2.10**, RNA ligation of a pre-adenylated 3' DNA Adapter (RA3, Illumina ref. # 15013207) was achieved first while using a truncated T4 RNA ligase 2 (NEB) (Supporting information *RNA ligation*). In the next step, we used the

SuperScript III® (SSIII) reverse transcriptase (RT) to synthesize the cDNA strands with a reverse primer (RA3, Illumina ref. # 15013981). SSIII RT is the third generation of engineered M-MLV RT. It has reduced RNase H activity and an increased stability at high temperature (up to 55ºC). After reverse transcription, all the samples are immediately loaded on an RNase-free agarose gel (**Figure 2.12**). We collect the gel in a range from 35bp to 300bp and extract the oligonucleotides with a "*crush-and-soak*" procedure (Supporting information, *Gel purification*). Given the size of the 3' adapter (21 nt), purified cDNAs correspond to transcripts from ~15 to ~280 nt long.

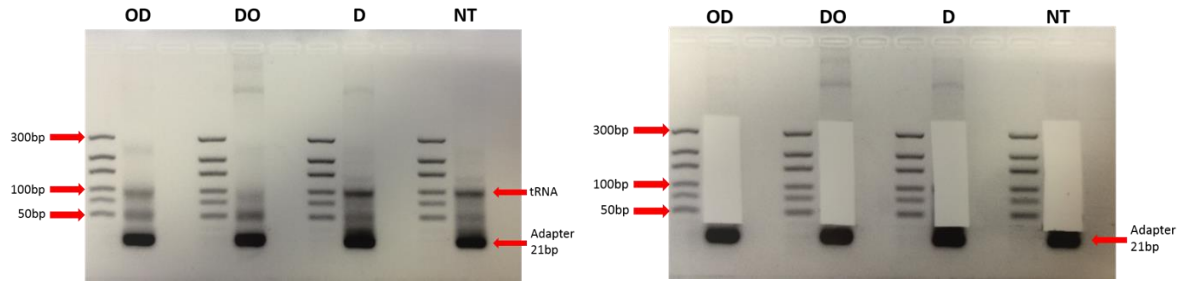



**Figure 2.12**. Gel purification of RT products of an $OD_{600}$=0.4 total RNA. Left: before gel extraction: two specific bands corresponding to tRNA and adapter are shown. Right: Same gel after extraction of the region of interest.

## Poly-A tailing

**T**erminal **d**eoxynucleotidyl **T**ransferase (TdT) is an uncommon DNA polymerase that adds nucleotides to the 3' terminus of DNA molecules without requiring any templates (Motea and Berdis 2010). Catalyzed by $Co^{2+}$, TdT can efficiently add 3'-end overhang to ssDNA, blunt end DNA or 3'-end recessed DNA in a few minutes. Taking of this advantage, TdT is widely used in modified mRNA 5' RACE and DNA library construction (Harvey and Darlison 1991; Lazinski and Camilli 2013). This enzyme can add any type of nucleotide. Because poly guanine oligo can format a G-quadruplexes (G4-DNA) and inhibit downstream PCR reaction, and could result in templates with high $T_m$ (up to 80ºC) together with Illumina RT primer, not compatible with PCR condition, we then choose to do a polyadenylation (Gellert et al. 1962; Rhodes and Lipps 2015; Tateishi-Karimata et al. 2016).

Following the guidelines of the TdT protocol (NEB, M0315S), the incubation conditions were adjusted so as to achieve the addition of ~10 to 30 adenosine residues: each total RNA

sample was incubated 45 min at 37ºC with 150 $\mu$M ATP (Supporting information, *TdT tailing*). **Figure 2.13** shows the analysis of the PCR products obtained after such TdT protocol.

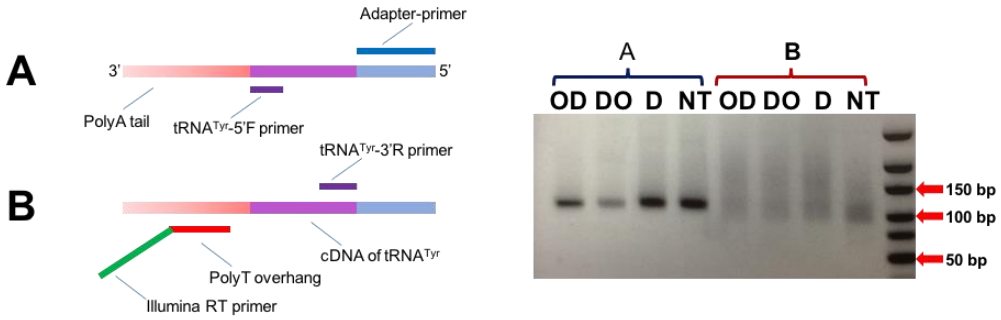



**Figure 2.13**. PCR control of cDNA polyadenylation by TdT. Group A PCR was achieved with $tRNA^{Tyr}$ 5' forward primer and primer of the adapter to amplify the cDNA generating from RNA ligation step. The expected size of PCR products is 107 bp including tRNA (85 nt) and adapter (22 nt). Group B PCR was achieved with $tRNA^{Tyr}$ 3' reverse primer and an Illumina RT primer with polyT overhang to amplify the polyandenylated cDNA. The expected size of PCR products should be more than 62 bp (corresponding to primer-dimmer).

We choose $tRNA^{Tyr}$ as a positive control to test the efficiency of polyadenylation. After polyadenylation and purification with Millipore 3K Spin column, 20 ng of treated cDNA from each sample was used run two PCR with different pairs of primers (indicated as A and B in **Figure 2.13**). In Group A PCR, the cDNA is amplified with a tRNA forward primer and the primer of the 3' adapter. The gel analysis shows that the positive selection (**OD**) has a stronger signal than the negative selection (**DO**), as expected. Furthermore, the other two controls **D** and **NT** both have a stronger signal than **OD** and **DO** samples. In Group B PCR, the cDNA is amplified with the tRNA reverse primer and the Illumina RT primer with $polyT_{17}$ overhang. All the samples show a faint smear, indicating that cDNA fragments with different sizes were amplified. The brighter signal present around 100 pb (stronger with the **NT** sample) suggests that reverse transcription significantly aborted after 30 nucleotides (total length $\approx$ Fwd. Adapt 46 nt + tRNA 30 nt + Rev. Adap 20 nt = 96 nt). A better overview of the poly-A tails characterizing these cDNA was provided by cloning and deep sequencing analyses (see below).

**PCR amplification**

A first PCR reaction of all samples (17 cycles with an annealing temperature gradient from 41ºC to 46ºC during the 5 first cycles; T(annealing) = 56ºC from cycle 6 to 17) was achieved to

add the forward universal primer while using the $dT_{17}$ overhang to prime on the added poly-A (Supporting information, *1r PCR amplification*). Gel Analysis revealed the presence of long smears, as expected (**Figure 2.14**). The excess of primers was removed from the PCR products with 30K Millipore spin columns.

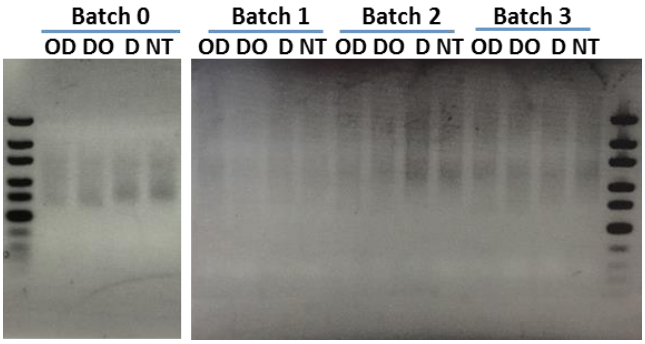



**Figure 2.14**. Gel analysis of the first PCR amplification with Illumina RT primer with $dT_{17}$ tail and primer of the adapter.

## Final PCR amplification

The second and final amplification allowed us to add the long Illumina primers to the samples, required for deep-sequencing multiplexing. The reverse universal primers are all characterized by a different index (**Table 2.6**).

**Table 2.6**. cDNA banks and Illumina Index Tag of the 16 samples.

| Sample ID | Insert size (bp) | Index Number | Index Sequence | Volume | Concentration |
|---|---|---|---|---|---|
| B0-1(OD) | 15-280 | Index 1 | CGTGAT | 20 $\mu$l | 40.5 ng/$\mu$l |
| B0-2(DO) | 15-280 | Index 2 | ACATCG | 20 $\mu$l | 35 ng/$\mu$l |
| B0-3(D) | 15-280 | Index 3 | GCCTAA | 20 $\mu$l | 56.8 ng/$\mu$l |
| B0-4(NT) | 15-280 | Index 4 | TGGTCA | 20 $\mu$l | 44.2 ng/$\mu$l |
| B1-1(OD) | 15-280 | Index 5 | CACTGT | 20 $\mu$l | 29.5 ng/$\mu$l |
| B1-2(DO) | 15-280 | Index 6 | ATTGGC | 20 $\mu$l | 40.8 ng/$\mu$l |
| B1-3(D) | 15-280 | Index 7 | GATCTG | 20 $\mu$l | 35.4 ng/$\mu$l |
| B1-4(NT) | 15-280 | Index 8 | TCAAGT | 20 $\mu$l | 27.2 ng/$\mu$l |
| B2-1(OD) | 15-280 | Index 9 | CTGATC | 20 $\mu$l | 38 ng/$\mu$l |
| B2-2(DO) | 15-280 | Index 10 | AAGCTA | 20 $\mu$l | 49.5 ng/$\mu$l |
| B2-3(D) | 15-280 | Index 11 | GTAGCC | 20 $\mu$l | 31.5 ng/$\mu$l |
| B2-4(NT) | 15-280 | Index 12 | TACAAG | 20 $\mu$l | 50.8 ng/$\mu$l |
| B3-1(OD) | 15-280 | Index 13 | TTGACT | 20 $\mu$l | 34 ng/$\mu$l |
| B3-2(DO) | 15-280 | Index 14 | GGAACT | 20 $\mu$l | 23.2 ng/$\mu$l |
| B3-3(D) | 15-280 | Index 15 | TGACAT | 20 $\mu$l | 31.6 ng/$\mu$l |
| B3-4(NT) | 15-280 | Index 16 | GGACGG | 20 $\mu$l | 21.5 ng/$\mu$l |

During final PCR, various amounts (50, 100 and 200 ng) of the first PCR were used to establish the best amplification conditions. Altogether 20 cycles of PCR were programmed to verify the presence of products. Gel analysis (**Figure 2.15**) showed that the majority of the products had a size significantly higher than adapter dimers, suggesting that most of the constructions had an insert.

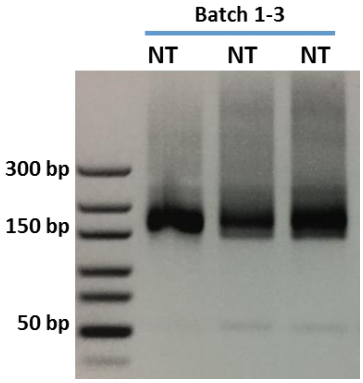



**Figure 2.15**. Gel analysis of a final PCR with 20 cycles (pre-test final cDNA bank, **NT** samples from batch 1 to 3)

In the final PCR, 100 ng of first round PCR product was loaded to do the second round 10 cycles PCR for each sample while using universal Illumina primer and different index primers. Gel analysis shows all of the visible products is a faint smear and the size is as expected (at least longer than 135 bp, corresponding to primer-dimer) (**Figure 2.16**, Left). Meanwhile, we also use the $tRNA^{Tyr}$ 5' forward primer and primer of the adapter to check all the samples before sending for sequencing. A significant difference between **OD** and **DO** treated sample is constantly observed (**Figure 2.14**, Right).

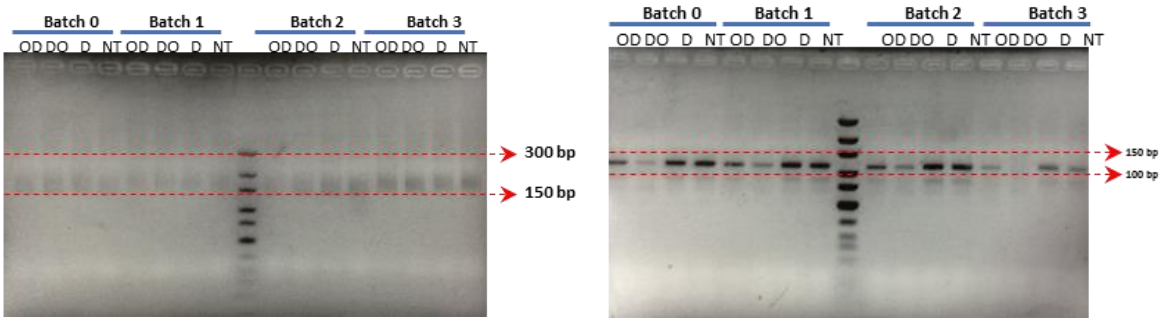



**Figure 2.16**. Left: Final PCR products size checking on 4% agarose gel. Right: PCR control with $tRNA^{Tyr}$.

### 2.2.2 Results

The samples were sent for sequencing (Beckman - Genewiz). An issue about the quality of these samples was revealed by bioanalyzer analyzes performed by them. There was a conflict about the size of the most abundant product: when our gel analysis suggested a size of at least 150 bp (**Figure 2.16**), the bioanalyzer traces showed a size of 135 bp, very close to the size of adapter dimer constructions (**Figure 2.17**).

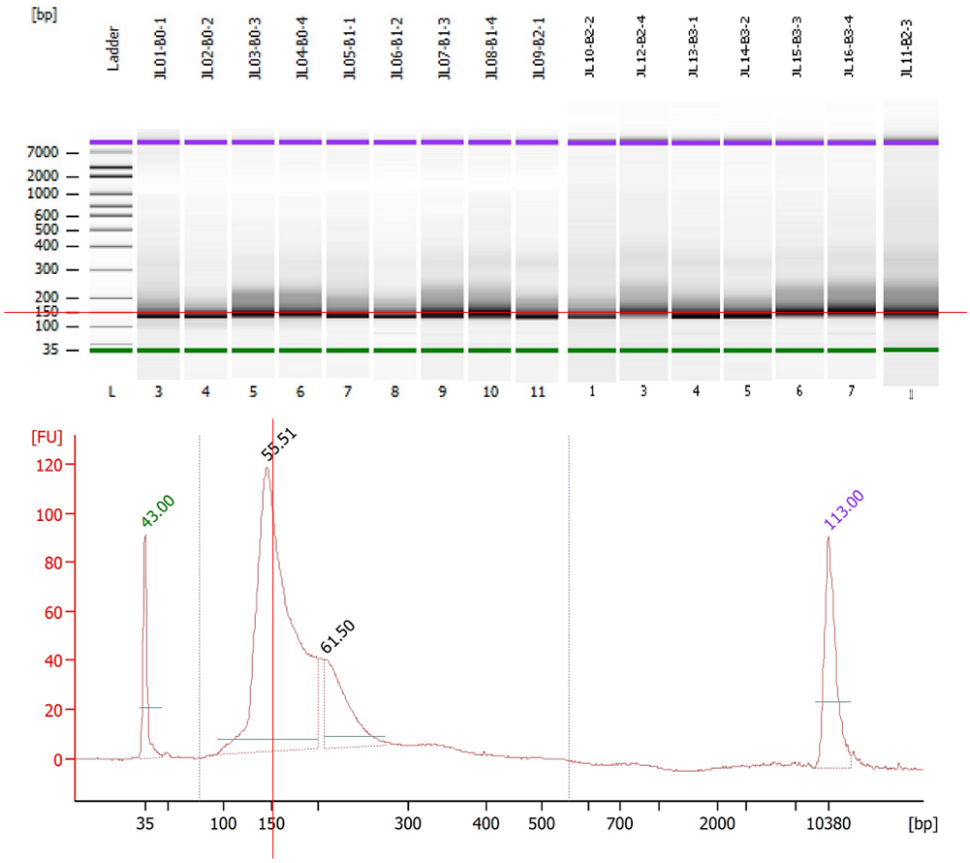



**Figure 2.17**. Bioanalyzer trace of a final PCR (final cDNA bank and one example of the **NT** sample)

To resolve this discrepancy, we decided to clone our samples and send colonies for sequencing. Among the "B-2" **OD** batch, only 4 out of 30 clones had an insert, two $tRNA^{Leu}$ fragments and two others unknown sequences (**Figure 2.18**), while the results where even worse (which is not surprised) with the "B-2" **DO** sample (2 out of 30 with an insert). Although this verification was disappointing, we still decided to proceed with deep sequencing, assuming

that even with 10% of constructions with inserts, the amount of information would be already high enough to possibly uncover some interesting data.

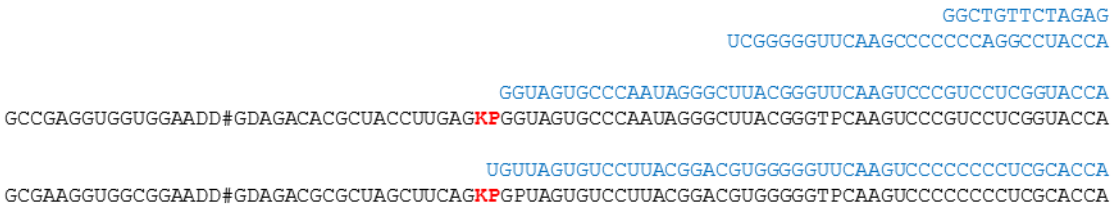



**Figure 2.18**. Single clone sequencing result of "B-2" **OD** sample. Blue shows the insert sequences. Black letter is the $tRNA^{Leu}$ full sequence and red letter indicates the base modification. #: 2'-*O*-methylguanosine; T: 5-methyluridine; D: dihydrouridine; P: pseudouridine; K: 1-methylguanosine. (http://modomics.genesilico.pl/)

## Deep-sequencing

As before, the 16 cDNA banks were processed by Beckman-Coulter USA (who became Genewiz in the meantime) for deep sequencing with the Illumina HiSeq2500 pair-end 2x125 technology.

It turned out that sequencing did not only provide very few reads (67 million altogether, i.e. only about 4 million reads per bank); the quality of these reads (FastQC report) was also extremely poor (% of >= Q30 bases: 9.79; Mean quality score: 14.66).

The origin of the low quality of the reads was quickly identified, and pointed out a flaw in the design of our forward universal primer. The first 4-5 sequencing cycles are devoted to the identification of cDNA clusters by the sequencer, which requires the highest possible diversity of sequences. Because the sequencer starts to collect information about the clusters at the position right after the forward universal primer, all the clusters gave a unique signal corresponding to a dT residue (from the $polyT_{17}$ overhang). The machine could thus not locate the precise position of these clusters, which had a dramatic effect on sequencing. The fact that the majority of the cDNA did not contain an insert also contributed to the very low overall quality of sequencing. The identification of this issue allowed us to design a new universal forward adapter that could eliminate the problem (see below).

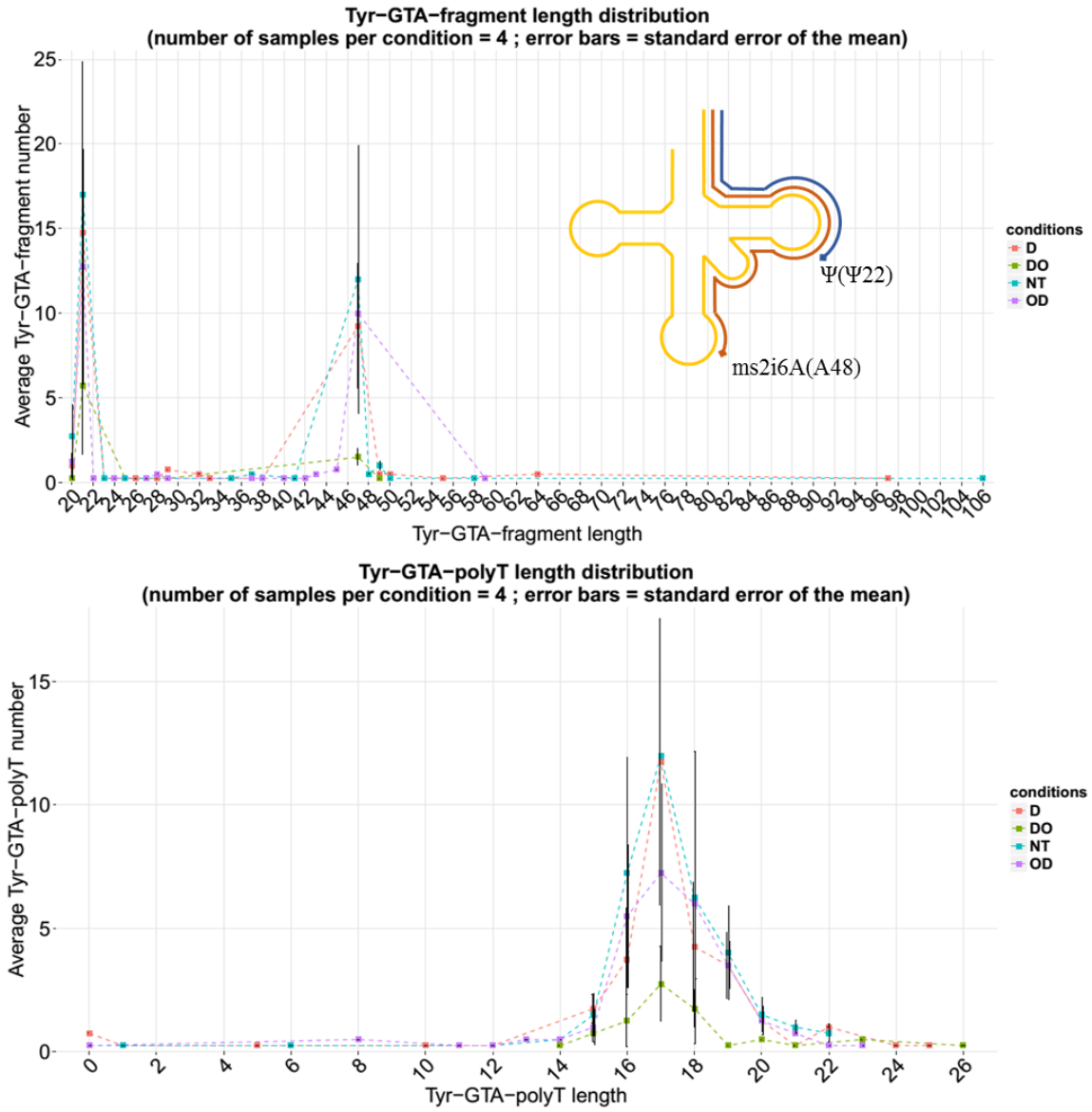
Despite the low number of reads and the high amount of sequencing errors, we still tried to extract some information from these reads while focusing on particular tRNA.

**Analysis of tRNA reads with OD, DO, D and NT samples**

An overview of the signals of four representative tRNA species (tyrosine- and lysine-tRNA) was achieved while collecting all the reads matching a 20-nt request corresponding to the first 20 nucleotides from the 3'-end of each tRNA. Because of the amount of sequencing errors, very few reads did match the request. This number could be increased while reducing the size of the request, consistent with the high number of errors.

Since the mapping of these reads on the *E. coli* genome could not provide any clear information, we decided to only collect some statistical properties from them. **Figure 2.19** presents the statistical distribution of the read lengths and the polyT lengths for each of the four tRNA.

**$tRNA^{Tyr}$ fragment. Average size: 47 nt and polyT length distribution**



**tRNA$^{Lys}$ fragment. Average size: 29 nt and polyT length distribution**

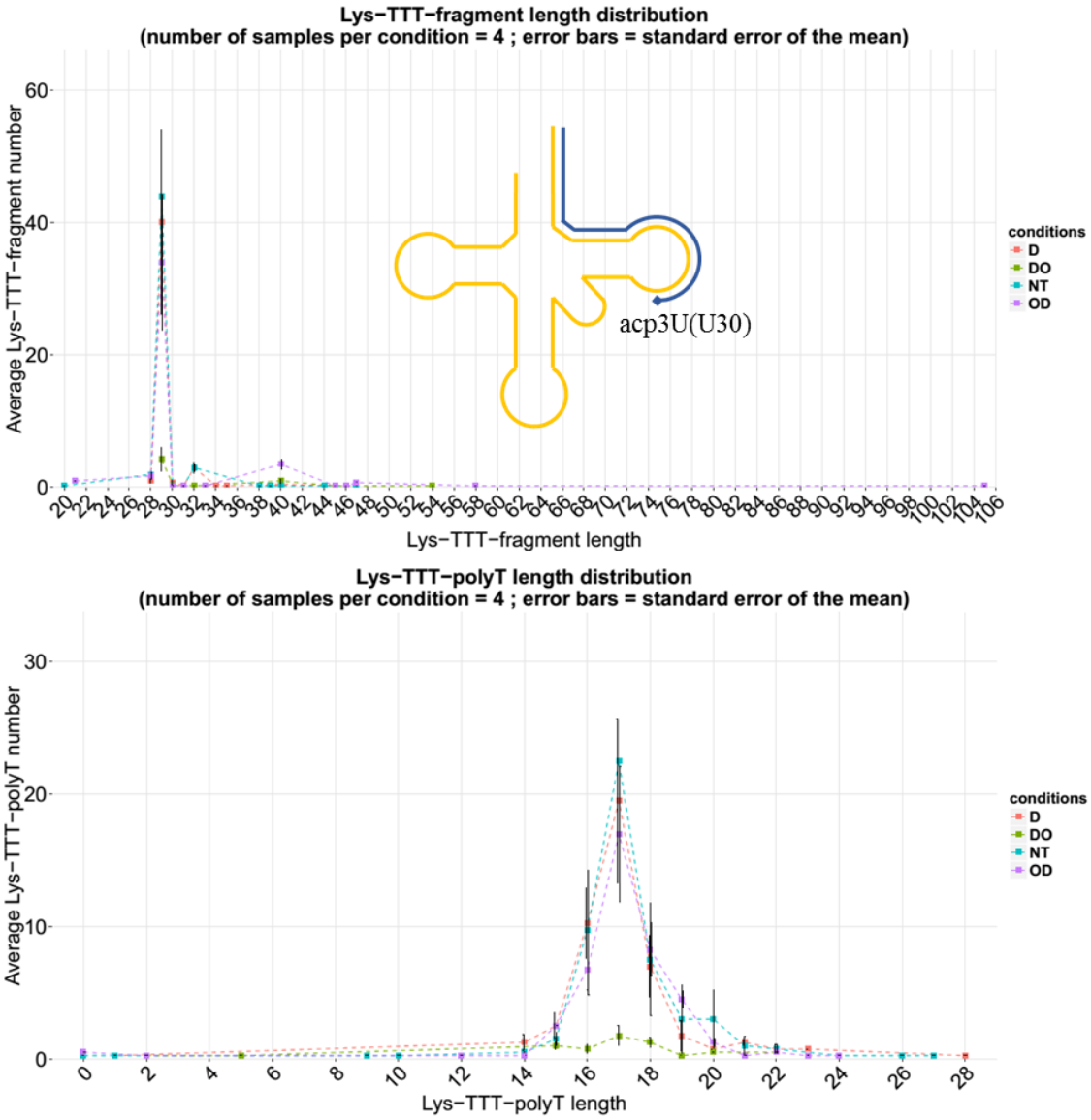



**Figure 2.19**. Distribution of tRNA$^{Tyr}$ and tRNA$^{Lys}$ reads length (up) and distribution of polyT length (down, both are 17 nt) for tRNA$^{Tyr}$ and tRNA$^{Lys}$ in each of the four **OD, DO, D, NT** conditions. In the figure, yellow cloverleaf shows the structure of full length tRNA. Brown or blue line shows the majority reads length alignment with the full length tRNA. The base modification is also indicated at *n*+1 position (Juhling et al. 2009).

Although the maximum number of reads is only of a few tens in each case, the observed distribution already reveals a few useful observations:

1 – It can be seen that DO samples have much fewer reads compared with the other three conditions, as expected.

2 – An examination of the tRNA$^{Tyr}$ and tRNA$^{Lys}$ reads shows the presence of a peak at 21, 47 nt (for tRNA$^{Tyr}$) and 29 nt (for tRNA$^{Lys}$), which correspond to base modification at that position (**Figure 2.19**). This result suggests that our sequencing strategy is promising.

### 2.2.3 Improved protocol

**New Illumina universal primer**

Newly designed forward adapter, with universal forward adaptor comprising a random NNNNN track at the end of the adapter, and before the polyT track which will allow sequencing machine to calibrate the signal at the beginning of sequencing process:


**PCR improvement to eliminate primer-dimer contamination**

In most of the sequencing bank preparation protocols, especially small RNA-seq, primer-dimer removal is a critical step. If the RNA in-put is rather low, both beads binding and gel purification are required (Quail et al. 2009; Kumar et al. 2012).

Primer-dimer contamination has also been revealed from our deep-sequencing results. Even though we did do the gel purification very carefully after reverse transcription, however, it is impossible to remove all the primer dimers. Possibly, total RNA could embed very few single stranded adapter and thus cause the downstream experiment generating primer-dimer by-product. From the PCR results we can see that this contamination is not a significant problem for **OD**, **D** and **NT** samples, while since the **DO** experiment is negative selection, the leftover RNA target will be much less than the primer-dimer which will cause the deep-sequencing data was dominated by the primer-dimer signal (**Figure 2.20**, left). This contamination will not only take up the quantity of meaningful reads, but also affect the spike normalization.

In order to avoid this contamination, we tried a new PCR strategy. Since we are interested in aminoacylated tRNA, we here decide to use a modified Illumina RT primer carrying three bases “TGG” at the 3’-end which is overhanging with the tRNA’s CCA-ending. With this new primer, primer-dimer is not visible even after 35 cycles PCR (**Figure 2.20**, right). Meanwhile, we also changed the spike to a fraction of human 18S RNA containing a CCA ending (5’-GACGAUCAGAUACCGUCGUAGUUCCGACCA-3’).

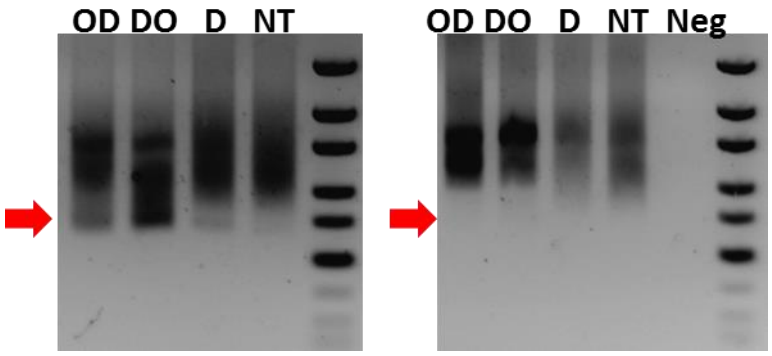
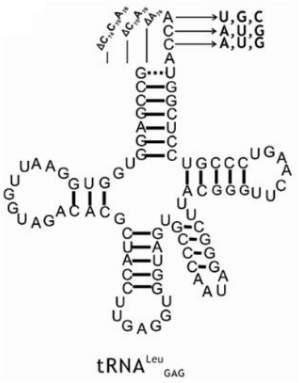



**Figure 2.20**. The prepared bank libraries are amplified by 35 cycles PCR *via* different primers. Left: the standard Illumina TruSeq small RNA library kit primers. Right: modified Illumina TruSeq small RNA library kit RT primer with "TGG" overhang. Red arrows point the position of primer-dimer (73 bp).

Obviously, the disadvantage of this modified PCR is to cause a globe bias that the final selected RNAs will all only have "CCA" ending. However, previous study had shown that the "CCA" ending is essential for the recognizing by aminoacyl tRNA synthetase (**Figure 2.21**). Base mutations will dramatically change the aminoacylation ration of tRNA (Zhou et al. 2011). In fact, the known tRNA-like RNAs so far all have similar tRNA motifs and some even highly mimic tRNA's 3'-end (See **Part II**). Here we assume that all the *in vivo* aminoacylated RNA should have a "CCA" ending.




| tRNA | $k_{cat}/K_m$ (relative) |
|---|---|
| $tRNA^{Leu}_{GAG}$ | 1 |
| A76U | 0.17 |
| A76G | – |
| A76C | 0.09 |
| C75A | 0.46 |
| C75U | 0.09 |
| C75G | 0.02 |
| C74A | 0.50 |
| C74U | 0.41 |
| C74G | 0.08 |

**Figure 2.21**. Aminoacylation kinetics of transcripts of $tRNA^{Leu}$ and $tRNA^{Leu}$ mutants from *E. coli*. (Zhou et al. 2011)

## Perspective

The first results obtained with our innovative protocol are not exploitable at this stage. However, several issues could be identified during this second attempt, and we are confident that with new samples prepared using an improved procedure, deep sequencing data will provide fully exploitable data with which exciting discoveries on tRNA aminoacylation, and possibly the discovery of new kind of aminoacylated RNA can be achieved.

# Supporting Information

## Experimental assay

## Chapter I

***DNA templates annealing***

T7 promoter (100 $\mu$M) 20 $\mu$l, cT7 DNA template (100 $\mu$M) 20 $\mu$l (N19 to N37 respectively), $H_2O$ 5 $\mu$l and NaCl (1 M) 5 $\mu$l; Mixture is incubated in 95ºC water bath for 5 min then cool down to room temperature overnight.

***RNA adapter annealing***

RNA adapter (Adapter Hyb-XL, N, NR, O, P) (100 $\mu$M) 1 $\mu$l, DEPC-water 99 $\mu$l; Heat the sample by PCR machine up to 90ºC for 2 min then drop the PCR tube into liquid nitrogen immediately, then do ethanol precipitation to collect the pellet for RNA ligation.

***Initial pool RNA T7 transcription and DNase treatment***

Annealed DNA templates (0.1 nmol), Ambion® T7 2X Buffer 10 $\mu$l, NTPs (10 mM) 8 $\mu$l, T7 RNA Polymerase 1 $\mu$l, $H_2O$ fills up to 20 $\mu$l; Mixture incubates at 37°C for 4 h. Add 2 $\mu$l Turbo DNase I (10 U/$\mu$l) and incubate 1h at 37°C. RNA products are purified by Phenol/Chloroform extraction.

***Phenol/Chloroform extraction***

Reaction solution 22 $\mu$l, NaOAc/HOAc buffer 15 $\mu$l, $H_2O$ 115 $\mu$l, Phenol/Chloroform/isoamyl alcohol (25:24:1) Mix (Sigma) pH 4-5 150 $\mu$l; Vortex 1 min and centrifuge for 10 min, then transfer the aqueous phase to a new tube for ethanol precipitation.

***Ethanol precipitation***

Aqueous phase 150 $\mu$l (including 15 $\mu$l NaOAc/HOAc buffer), Glycogen (ThermoFisher) 1 $\mu$l and 100% ethanol 450 $\mu$l. Incubate at -20ºC for at least 45 min; Centrifuge sample for 30 min at 4ºC with speed 15000 rpm then wash the pellet by 70% cold ethanol (DO NOT vortex), continuously centrifuge for another 15 min at 4ºC with the same speed. Air dry the pellet and re-suspend in DEPC-water.

***The 1st cycle selection***

1) ***Preparation***

Prepare sterilized 1 M NaCl, KCl, NaOH solution, 100 mM ATP and GTP (ThermoFisher), 250 mM $CaCl_2$, 250 $\mu$M $MnCl_2$ solution. Dilute 1M $MgCl_2$ (Sigma) to 250 mM. Prepare 0.01% poly-lysine and poly-arginine as well as 250 $\mu$M spermine (Sigma). Amino acids mix (Sigma) contains 2.5 mM Ala, Arg, Asp, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Tyr, Val and 1.25 mM Cys.

2) ***Aminoacylation buffer***

| **Components** | **Volume** | **Final concentration** |
|---|---|---|
| RNA (around 5 $\mu$M) | (adjustable) 5 $\mu$l | 0.25 $\mu$M (1-3cycle), 0.1 $\mu$M (remaining cycles) |
| NaCl (1 M) | 10 $\mu$l | 100 mM |
| KCl (1 M) | 10 $\mu$l | 100 mM |
| Total | $H_2O$ fills up to 30 $\mu$l | |
| Incubate at 90°C for 3 min and cool down to 60°C | | |
| $CaCl_2$ (250 mM) | 1 $\mu$l | 2.5 mM |
| $MgCl_2$ (250 mM) | 2 $\mu$l | 5 mM |
| $MnCl_2$ (250 $\mu$M) | 2 $\mu$l | 5 $\mu$M |
| Poly-lysine (0.01%) | 2 $\mu$l | |
| Poly-arginine (0.01%) | 2 $\mu$l | |
| Spermine (250 $\mu$M) | 1 $\mu$l | 2.5 $\mu$M |
| aa Mix (42.5 mM) | 40 $\mu$l | 17 mM |
| Mix well and chill on ice | | |
| GTP (100 mM) | 2 $\mu$l | 2 mM |
| ATP (100 mM) | 2 $\mu$l | 2 mM |
| Mix well and avoid any precipitation | | |
| NaOH (1 M) | 3.5 $\mu$l | 35 mM<br>pH = 3.5 |
| Total | 88 $\mu$l | |
| Incubate on ice overnight | | |
| HEPES buffer (1 M, pH=7.0) | 12.5 $\mu$l | pH 7.0 |
| Total | 100.5 $\mu$l | |
| Incubate on ice for 2h and then precipitate | | |

3) ***OD-DO Assay***

Cover all the tubes with aluminum films; Dissolved 5.4 mg $NaIO_4$ in 250 $\mu$l water to prepare 0.1 M sodium-periodate solution; All the oxidations are performed in dark and on ice. After each oxidation, add 10 $\mu$l Glucose (0.5 M) and equilibrate for 15 min on ice. Then add 6 $\mu$l DTT before precipitation.

| | OD | DO | D | NT |
|---|---|---|---|---|
| NaOAc/HOAc (3M pH=5.2) | 5 $\mu$l | 0 | 0 | |
| $H_2O$ | Up to 50 $\mu$l | Up to 50 $\mu$l | Up to 50 $\mu$l | Up to 50 $\mu$l |
| $NaIO_4$ sol. (0.1M) | 5 $\mu$l | 0 | 0 | |
| Borax buffer (pH=10.0) | 0 | 7 $\mu$l | 7 $\mu$l | |
| | on ice 45 min | 42ºC 1h | 42ºC 1h | on ice 1h |
| | | Precipitation | | |
| NaOAc/HOAc | 0 | 5 $\mu$l | | |
| $H_2O$ | Up to 50 $\mu$l | Up to 50 $\mu$l | | |
| $NaIO_4$ sol. (0.5M) | 0 | 5 $\mu$l | | |
| Borax buffer (pH=10) | 7 $\mu$l | 0 | | |
| | 42ºC 1h | on ice 45 min | | |
| | | Precipitation | | |

***T4 RNA ligation***

RNA pellet, Adaptor pellet (0.1 nmol), $H_2O$ 14 $\mu$l, 10X T4 RNA Ligase Buffer 4 $\mu$l, T4 RNA Ligase (5 U/$\mu$l) 2 $\mu$l (ThermoFisher), PEG8000 (50%) 20 $\mu$l; Mix well and incubate at 10°C overnight. Then precipitate to remove the reaction buffer.

***Reverse transcription***

RNA ligation pellet, $H_2O$ 13 $\mu$l, dNTPs (10 mM) 1 $\mu$l, DTT (0.1 M) 1 $\mu$l, 5X RT Buffer 4 $\mu$l, SuperScript® III (200 U/$\mu$l) 0.4 $\mu$l; Incubate at 50 °C for 30 min then precipitation. Re-suspend the pellet in 20 $\mu$l DEPC-water.

***(T4 DNA ligation)***

Reverse transcription pellet, $H_2O$ 15 $\mu$l, DNA Anchor (10 $\mu$M) 1 $\mu$l, 10X T4 DNA ligation buffer 2 $\mu$l, T4 DNA ligase (10 U/$\mu$l) 2 $\mu$l. Incubate at r.t. overnight. Then precipitate, re-suspend into 20 $\mu$l water.

***Short primer PCR***

Prepare the PCR mix: $H_2O$ 63 $\mu$l, cDNA template 4 $\mu$l, Adaptor Primer P + TGA (10 $\mu$M) 8 $\mu$l, T7promoter + CACGA (10 $\mu$M) 8 $\mu$l, dNTPs (5 mM) 4 $\mu$l, 10X DreamTaq Green Buffer 10 $\mu$l, DreamTaq DNA Polymerase (1 U/$\mu$l) 0.4 $\mu$l. Split into two tubes (50 $\mu$l/tube). Run 35 cycles PCR (94°C, 2 min; [94°C 30s, 60°C 45s, 72°C 20s]x35; 72°C 2 min).

***Gel purification***

Use 30K Amicon® Ultra-0.5 Centrifugal Filter Millipore spin column to concentrate 5 tubes PCR products and load on 4% agarose gel with ethidium bromide. Run the gel at 135 voltages for 30 min in 0.5X TAE electrophoresis running buffer. Cut the gel at expected size and chop it into small pieces. Soak the gel into 450 $\mu$l $H_2O$ + 50 $\mu$l TE Buffer (Sigma) and shake at 20ºC overnight. Precipitate to collect DNA pellet.

***Re-amplification***

PCR mix: $H_2O$ up to 100 $\mu$l, DNA template 10 ng, Adaptor Primer P + TGA (10 $\mu$M) 4 $\mu$l, T7promoter + CACGA (10 $\mu$M) 4 $\mu$l, 10X AccuPrime™ *Taq* Buffer I (including $Mg^{2+}$ and dNTPs) 10 $\mu$l, AccuPrime™ *Taq* hot-start DNA Polymerase (1 U/$\mu$l) 0.4 $\mu$l. Split into two tubes (50 $\mu$l/tube). Run 20 cycles PCR (94°C, 2 min; [94°C 30s, 60°C 45s, 72°C 20s]x20; 72°C 2 min).

***EarI (NEB) digestion***

Load 100 $\mu$l PCR products on to SigmaSpin™ Sequencing Reaction Clean-Up Columns to remove the reaction buffer and precipitate to collect the DNA pellets. DNA 1 µg, restriction enzyme (20 U/$\mu$l) 1$\mu$l, 10X NEB buffer 5 $\mu$l, $H_2O$ fills up to 50 $\mu$l. Incubate at 37ºC overnight. Clean the digested products by Phenol/Chloroform extraction.

***<u>The 2nd round T7 transcription</u>***

0.1 nmol digested PCR products are used as template for the next round RNA synthesis.

***The 4th and remaining rounds transcripts HPLC purification***

HPLC column: Waters XTerra™ MS $C_{18}$ 2.5 m (4.6 mm x 50 mm) SN:01713519613005; Working temperature: 50ºC; Solution A: 0.1 M 95% Triethylammonium acetate (TEAA) and 5% Acetonitrile (ACN). Solution B: 0.1 M 85% TEAA and 15% ACN. The gradient of A% is 80% to 10% in 20 min (0.9 ml/min). Retention time for N19 to N37 library is 5, 6, 7.5 and 8 min respectively. Collected RNA is concentrated by 3K Amicon® Ultra-0.5 Centrifugal Filter and precipitate for the next round selection.

***Bioanalyzer***

Follow the standard protocol of Small RNA Chip.

***Aminoacylation verified by HPLC***

HPLC column: Waters XTerra™ MS $C_{18}$ 2.5 m (4.6 mm x 50 mm) SN:01713519613005; Working temperature: 50ºC; Solution A: 0.1 M 95% Triethylammonium acetate (TEAA) and 5% Acetonitrile (ACN). Solution B: 0.1 M 85% TEAA and 15% ACN. The gradient of A% is 80% to 45% in 25 min (0.9 ml/min).

# Chapter II

***Total RNA extraction***

All the experiments have to be done with gloves, on ice and manipulated with RNase-free tips and DEPC-water. *E. coli* strain either BL21 (PhB101) or MG1655 (PhB1463) is from the lab collection of Philippe Bouloc. Streak the *E. coli* strain on LB plate and incubate at 37ºC overnight to pick a single clone for pre-culture. Run LB pre-culture at 37ºC overnight (180 rpm). Prepare Falcon tube with either 7ml or 15 ml 100% Ethanol at -80ºC overnight. The next day, run three parallel cultures at the same time. Dilute 1 ml pre-culture into 100 ml fresh LB culture and incubate at 37ºC (180 rpm). Collect three $OD_{600}$ points at 0.2, 0.4 and 0.8 culture to quench in equal volume of 15 ml cold ethanol ($OD_{600}$ 0.2) or 7 ml cold ethanol ($OD_{600}$ 0.4 and 0.8). Centrifuge Falcon tube 15 min (4000 rpm) at 4ºC and re-suspend the pellet in 600 $\mu$l lysis buffer (0.02 M NaOAc/HOAc pH 5.5, 0.5% SDS, 1 mM EDTA). Mix the lysate pellet with 600 $\mu$l acid phenol and 100 $\mu$l water, vortex to mix well. Add 1/10 volume 3M NaOAc/HOAc pH 5.5 and vortex to mix well. Centrifuge 5 min at 13000 *g* and transfer the aqueous phase to a new tube. Repeat the Phenol extraction step. Transfer the aqueous phase into a new tube and mix with 600 $\mu$l chloroform. Centrifuge and collect the aqueous phase to do ethanol precipitation overnight. Re-suspend the total RNA pellet into 100 $\mu$l DEPC-water and aliquot in several PCR tubes storing at -80ºC.

***OD-DO treatment***

Cover all the tubes with aluminum films; Dissolved 5.4 mg $NaIO_4$ in 250 $\mu$l water to prepare 0.1 M sodium-periodate solution; All the oxidations are performed in dark and on ice. After each oxidation, add 10 $\mu$l Glucose (0.5 M) and equilibrate for 15 min at room temperature. Then add 6 $\mu$l DTT before precipitation.

| | OD | DO | D | NT |
|---|---|---|---|---|
| | 4 $\mu$g total RNA for each sample | | | |
| NaOAc/HOAc (3M pH=5.2) | 5 $\mu$l | 0 | 0 | |
| $H_2O$ | Up to 50 $\mu$l | Up to 50 $\mu$l | Up to 50 $\mu$l | Up to 50 $\mu$l |
| DMSO | 2 $\mu$l | 2 $\mu$l | 2 $\mu$l | 2 $\mu$l |
| $NaIO_4$ sol. (0.1M) | 5 $\mu$l | 0 | 0 | |
| Borax buffer (pH=9.0) | 0 | 7 $\mu$l | 7 $\mu$l | |
| | r.t. 30 min | 42ºC 45 min | 42ºC 45 min | on ice 45 min |
| | Precipitation | | | |
| NaOAc/HOAc | 0 | 5 $\mu$l | | |
| $H_2O$ | Up to 50 $\mu$l | Up to 50 $\mu$l | | |
| DMSO | 2 $\mu$l | 2 $\mu$l | | |
| $NaIO_4$ sol. (0.5M) | 0 | 5 $\mu$l | | |
| Borax buffer (pH=9.0) | 7 $\mu$l | 0 | | |
| | 42ºC 45 min | r.t. 30 min | | |
| | Precipitation | | | |

***DNase treatment and spike adding (1st sequencing)***

Re-suspend treated RNA (around 1~4 $\mu$g) pellet in 88 $\mu$l $H_2O$, 10X buffer Turbo® DNase 10 $\mu$l, Turbo DNase (U/$\mu$l) 2 $\mu$l; Incubate 1h at 37°C then keep at 16°C overnight. Do Phenol/Chloroform extraction the next day. Measure the concentration of all the treated samples OD/DO/D/NT by Nano-drop. Add exactly the same mass percentage pre-mixed Spike-25/65/135 nt into 4 samples, 5% per sample. After spike adding, verify the final concentration. Run bioanalyzer to check the RNA quality. PCR to verify the DNase treatment efficiency by using the primers pairs tRNA-tyr5'/tRNA-tyr3'. PCR mix: RNA sample 1 $\mu$l, DreamTaq polymerase (ThermoFisher) 0.1 $\mu$l, 10X DreamTaq Green buffer 2.5 $\mu$l, dNTPs (5mM) 2 $\mu$l, primer Tyr forward and reverse (10 $\mu$M) 2 $\mu$l for each, $H_2O$ fills up to 25 $\mu$l. Run 30 cycles PCR (95°C 2 min; [95°C 20s, 55°C 20s, 72°C 20 s]; 72°C 2 min)

***DNase treatment and spike adding (2nd sequencing)***

Re-suspend treated RNA (around 1~4$\mu$g) pellet in 88 $\mu$l $H_2O$, 10X buffer Turbo® DNase 10 $\mu$l, Turbo DNase (U/$\mu$l) 2 $\mu$l; Incubate 4h at 37°C then keep at 16°C overnight. Do Phenol/Chloroform extraction the next day. Measure the concentration of all the treated samples OD/DO/D/NT by Nano-drop. Add exactly the same mass percentage Spike-65nt into 4 samples, 2% per sample. After adding spike, verify the final concentration. Run 4% agarose gel at 135 voltages for 35 min to check the RNA quality. PCR to verify the DNase treatment efficiency by using the primers pairs tRNA-tyr5'/tRNA-tyr3'. PCR mix: RNA sample 1 $\mu$l, DreamTaq polymerase (ThermoFisher) 0.1 $\mu$l, 10X DreamTaq Green buffer 2.5 $\mu$l, dNTPs (5mM) 2 $\mu$l, primer Tyr forward and reverse (10 $\mu$M) 2 $\mu$l for each, $H_2O$ fills up to 25 $\mu$l. Run 30 cycles PCR (95°C 2 min; [95°C 20s, 55°C 20s, 72°C 20 s]x30; 72°C 2 min)

***RNA ligation***

OD, DO, D, NT treated RNA (1 $\mu$g per tube) 5 $\mu$l (adjustable), Pre-adenylated RNA 3' adapter (50 $\mu$M) 1 $\mu$l; Mix well and then place on PCR machine, denature at 70°C for 2 min then chill on ice immediately. Add 10x T4 RNA ligase 2 buffer 1 $\mu$l, DMSO 1 $\mu$l, BSA 1 $\mu$l, T4 RNA ligase 2 truncated (NEB) 1 $\mu$l. Incubate at 10°C overnight. The next day add 40 $\mu$l $H_2O$ then precipitation. After adding the ethanol, add 1 $\mu$l RT primer (50 $\mu$M) in the solution. After precipitation, re-suspend the pellet into 13 $\mu$l DEPC-water.

***Reverse transcription***

RNA ligation pellet 13 $\mu$l, dNTPs mix (10 mM) 1 $\mu$l; Incubate at 70°C for 2 min then chill on ice. Add 5X RT Buffer 4 $\mu$l, DTT (100 mM) 1 $\mu$l, SuperScript III® Reverse transcriptase (200 U/$\mu$l) 1 $\mu$l. Incubate 1 h at 55°C then chill on ice and ready to load on gel

***Agarose gel purification***

Prepare 3% agarose gel with ethidium bromide (2.4 g agarose and 80 ml TAE 0.5X running buffer). Load the reverse transcription products directly on gel. Run at 135 voltages for 5 min, then switch to 100 voltages for 75 min. Cut the gel from 35 bp to 200 bp.

***Break the gel and recover cDNA***

Prepare the gel breaker tubes by using 0.8 mm needle punch the bottom of 0.5 ml tube 3-4 times to create small holes. Cut the gel into suitable size (DO NOT chop into small pieces) and put it into

the 0.5 ml gel breaker tube, then cover it with a 2 ml tube. Centrifuge at 15000 rpm for 5-10 min to make sure there is no gel left. Suspend crushed gel by 450 $\mu$l $H_2O$ and 50 $\mu$l TE buffer per tube then shake overnight at 20°C. The next day, transfer gel solution into a Corning Costar Spin-X® filter tube then centrifuge at 15000 rpm for 15 min, collect the wash flow and precipitation.

***TdT tailing***

Recovered cDNA 500 ng, 5X TdT buffer 10 $\mu$l, 10 mM dATP 0.75 $\mu$l, TdT enzyme (20U/$\mu$l) (ThermoFisher) 0.5$\mu$l, $H_2O$ fills up 50 $\mu$l. Incubate at 37°C for 45 min then stop the reaction by heating up to 70°C for 10 min (10-30 A will be added). Using Millipore 10K spin column clean the reaction solution (50 $\mu$l reaction solution + 450 $\mu$l water, spin 10 min; refill 450 $\mu$l water, spin 20 min, collect leftover). Precipitate overnight and suspend the pellet into 45 $\mu$l $H_2O$ and 5 $\mu$l TE buffer.

***The first round PCR***

PCR Mix: $H_2O$ 37 $\mu$l, 10X AccuPrime™ buffer I 5 $\mu$l, Illumina RNA RT primer (10 $\mu$M) 1 $\mu$l, Reverse polyT primer (10 $\mu$M) 1 $\mu$l, cDNA template (50 ng) 5 $\mu$L, AccuPrime™ Taq polymerase 1 $\mu$l; Run 10 to 18 cycles (depends) PCR; (94°C 2 min; [94°C 30 s, 41+1°C 40 s, 68°C 20 s] x 5; [94°C 30 s, 56°C 30 s, 68°C 20 s] x 13; 68°C 2 min).

***PCR products purification***

Using 30K Amicon® Ultra-0.5 Centrifugal Filter Millipore spin column to clean the reaction solution (50 $\mu$l reaction solution + 450 $\mu$l water, spin 15 min; refill 450 $\mu$l water, spin 15 min, collect leftover). Precipitate overnight and suspend the pellet into 20 $\mu$l $H_2O$

***The second round PCR and Index adding***

PCR mix: 10X AccuPrime buffer II 5 $\mu$l, Universal Primer (10 $\mu$M) 1 $\mu$l, Illumina Index Tag 1~16 (10 $\mu$M) 1 $\mu$l, Pre-PCR Template (1~16) 100 ng, AccuPrime Taq polymerase 1 $\mu$l, $H_2O$ fills up to 50 $\mu$l. (94°C 2 min; [94°C 30 s, 60°C 40 s, 68°C 20 s] x 10; 68°C 2 min)

***Gel purification***

Using 30K Amicon® Ultra-0.5 Centrifugal Filter Millipore spin column to clean the reaction solution (50 $\mu$l reaction solution + 450 $\mu$l water, spin 15 min; refill 450 $\mu$l water, spin 15 min, collect leftover). Prepare 1.2% agarose gel with ethidium bromide. Load PCR products and run the gel with fresh 0.5X TAE buffer at 135 voltages for 25 min. Cut the gel from 150 bp to 400 bp and break the gel by gel breaker tube. Suspend crushed gel into 450 $\mu$l water + 50 $\mu$l TE buffer and shake at 37°C overnight. Precipitation and re-suspend sample in pure water and verify the concentration.

# Oligo sequences

| **Templates Sequence (5'→3')** | |
|---|---|
| **N19** | YGANNNNNNNNNTCGTGCCTATAGTGAGTCGTATTAGGATCC |
| **N23** | YGANNNNNNNNNNNNNTCGTGCCTATAGTGAGTCGTATTAGGATCC |
| **N30** | YGANNNNNNNNNNNNNNNNNNNNTCGTGCCTATAGTGAGTCGTATTAGGATCC |
| **N37** | YGANNNNNNNNNNNNNNNNNNNNNNNNNNNTCGTGCCTATAGTGAGTCGTATTAGGATCC |
| **Adapter (5'→3')** | |
| Adapter K | /5Phos/rUrUrUrUrGrArArGrArGrCrGrGrCrCrG |
| Adapter L | /5Phos/rUrUrUrUrCrA**rGrGrArUrArC**rGrCrCrGrCrU/3InvdT/ |
| Adapter XL | /5Phos/rUrUrUrUrCrA**rGrGrArUrArC**rGrGrArGrUrUrArArCrUrUrUrGrCrArUrArGrGrCrGrArUrUrGrCrArA/3InvdT/ |
| Hyb-Adapter XL | /5Phos/rUrUrUrUrCrA**rGrGrArUrArC**rUrCrGrCrArGrUrUrArArCrUrCrGrGrCrArUrArGrGrCTTTTGCCTATGCCGAGTTAACTGC |
| Adapter M | /5Phos/rUrUrUrUrCrA**rGrGrArUrArC**rGrCrArGrUrCrUrArCrUrG/3InvdT/ |
| Adapter N | /5Phos/rUrUrUrUrGrU**rGrGrArUrArC**rGrCrArGrUrCrUrArCrUrGTTTTCAGTAGACTGCGTATCCACAAAATGG |
| Adapter NR | /5Phos/rUrUrUrUrGrU**rGrGrArUrArC**rGrCrArGrUrCrUrArCrUrGTTTTCAGTAGACTGCGTATCCACAAAATGGR |
| Adapter O | /5Phos/rUrUrUrUrU**rCrCrCrArGrU**rCrArGrCrUrGrUrCrUrArGrUTTTTACTAGACAGCTGACTGGG |
| Adapter P | /5Phos/rA**rGrArArGrArG**rCrCrGrUrUrArGrCrUrGrUrCrUrArGrUTTTTACTAGACAGCTAACGGCTC |
| **Primer of the adapter** | |
| Primer K | 5-CGGCCGCTCTTCAAAA-3 |
| Primer L | 5-TTGCAATCGCCTATGCAA-3 |
| Primer XL+TGC | 5-AAGTTAACTCCGTATCCTGAAAATGC-3 |
| Primer XL32 | 5-TTGCAATCGCCTATGCAAAGTTAACTCCGTAT-3 |
| Hyb-Primer XL | 5-GCCTATGCCGAGTTAACTGC-3 |
| Primer M | 5-CAGTAGACTGCGTATCCTGAAAA-3 |
| Primer M+TGG | 5-CAGTAGACTGCGTATCCTGAAAATGG-3 |
| Primer N+TGG | 5-CAGTAGACTGCGTATCCACAAAATGG-3 |
| Primer O | 5-ACTAGACAGCTGACTGGGAAAAA-3 |
| Primer O+TGA | 5-TTTTACTAGACAGCTGACTGGGAAAAATGA-3 |
| Primer O+TGG | 5-ACTAGACAGCTGACTGGGAAAAATGG-3 |
| Primer O+CGA | 5-ACTAGACAGCTGACTGGGAAAAACGA-3 |
| Primer O-XX-TGA | 5-TTTTACTAGACAGCTGACTGGGAAA/idSp//idSp/TGA-3 |
| Primer P+TGA | 5-TTTTACTAGACAGCTAACGGCTCTTCTTGA-3 |
| **mini-tRNA control sequences (5'→3')** | |
| miniTRNA-VAL(TAC)5'-51b | rGrGrUrUrCrGrArArCrCrCrGrUrCrArUrUrCrUrCrCrArCrCrA |

| | |
|---|---|
| miniTRNA-VAL(TAC)5'-25b | rArUrCrUrGrCrCrUrUrArCrArArGrCrArGrArGrGrUrCrGrGr CrGrGrUrUrCrGrArArCrCrCrGrUrCrArUrUrCrUrCrCrArCrCrA |
| **RNA Spikes (5'→3')** | |
| RNA Spike_25 | rGrGrGrArGrArGrArGrGrGrCrUrGrCrUrGrUrUrCrUrArGrArG |
| RNA Spike_65 | rGrGrGrArGrArGrArGrArGrArGrArArUrUrArCrCrCrUrCrUrArCrGrCrUrA rUrUrGrGrArGrCrUrGrGrArArUrUrUrCrCrGrCrGrGrCrUrGrCrUrGrUrC rUrArGrArG |
| RNA Spike_135 | rGrGrGrArGrArGrArGrArGrArGrArGrArArUrUrArCrCrCrUrCrArCrUrArArArG rGrGrArGrGrArGrArArGrCrUrUrArUrCrCrCrArArGrArUrCrCrArArCrUrA rCrGrArGrCrUrUrUrUrUrArArCrUrGrCrArGrCrArArCrUrUrUrArArUrArU rArCrGrCrUrArUrUrGrGrArGrCrUrGrGrArArUrUrUrCrCrGrCrGrGrCrUrG rCrUrGrUrUrCrUrArGrArG |
| **Primer pairs** | |
| Saur_tRNAval3 | 5-TGGTGGAGAATGACGGGT-3 |
| Saur_tRNAval5(+51) | 5-GGTTCGAACCCGTCATTCT-3 |
| Saur_tRNAval5(+25) | 5-ATCTGCCTTACAAGCAGAGGG-3 |
| Saur_tRNAala3 | 5-TGGTGGAGACTAGCGGGA-3 |
| Saur_tRNAval5(+23) | 5-CTGCTTTGCACGCAGGA-3 |
| Saur_tRNAgly3 | 5-TGGAGCAGAAGACGGGAT-3 |
| Saur_tRNAgly5(+25) | 5-ACAACCTTGCCAAGGTTGG-3 |
| Ecoli BL21(tTyr)-R3' | 5-TGGTGGTGGGGGAAG-3 |
| Ecoli BL21(tTyr)-F5' | 5-GGTGGGGTTCCCGAG-3 |
| Ecoli BL21(tVal-26)-F5' | 5-CCTCCCTTACAAGGAGGG-3 |
| Ecoli BL21(tVal)-R3' | 5-TGGTGGGTGATGACGGGA-3 |
| Ecoli BL21(tAla-26)-F5' | 5-CCTGCTTTGCACGCAGGA-3 |
| Ecoli BL21(tAla)-R3' | 5-TGGTGGAGCTATGCGGGA-3 |
| **Short primer PCR test oligo (5'→3')** | |
| 9ntSpec_PCR_Ctrl | ACTAGACAGCTGACTGGGAAAAATGAN(25)ATTCGTGCC |
| T7Promoter_7nt_PosiCtrl | GGATCCTAATACGACTCACTATAGGCACGA |
| T7Promoter_8nt_PosiCtrl | GGATCCTAATACGACTCACTATAGGCACGAA |
| T7Promoter_9nt_PosiCtrl | GGATCCTAATACGACTCACTATAGGCACGAAT |
| **DNA anchor up strand (cT7 Promoter ligase) (5'→3')** | |
| cT7Prom-T7Prom12GGN-PEG | /5Phos/TATAGTGAGTCGTATTAGGATCC/iSp9/ACTCACTATAGGN/3Sp9/ |
| cT7PromoterLigase | /5Phos/TATAGTGAGTCGTATTAGGATCC |
| cT7_Promoter (32)_SpC3 | /5Phos/TATAGTGAGTCGTATTAGGATCCGGCGG*G*C/3SpC3/ |
| cT7_CC_Promoter (32)_SpC3 | /5Phos/CCTATAGTGAGTCGTATTAGGATCCGGCGG*G*C/3SpC3/ |
| **DNA anchor down strand (XL T7 Promoter) (5'→3')** | |
| XLT7_Promoter_3N_12I | GCCGGATCCTAATACGACTCACTATAGGNNNIIIIIIIII*I*I/3SpC3/ |
| XLT7_Promoter2N10I | CCGCCGGATCCTAATACGACTCACTATAGGNNIIIIIIII*I*I/3SpC3/ |
| XLT7_Promoter8N | CCGCCGGATCCTAATACGACTCACTATAGGNNNNNN*N*N/3SpC3/ |
| XLT7_PromoterCA4I | CCGCCGGATCCTAATACGACTCACTATAGGCAII*I*I/3SpC3/ |
| **Tested sequences (5'→3')** | |

| | |
|---|---|
| JA1 | GGCACGAUGCCGCGGGUUCGGGCGCUAUCA |
| JA2 | GGCACGAUCAUGUCA |
| JA3 | GGCACGACCCGCGCCUCA |
| JA4 | GGCACGACAGCACGCCUCA |
| JA5 | GGCACGACAGCCCCGCCUCA |
| JA6 | GGCACGAACAGCUGCCCGUCUCA |
| JA7 | GGCACGACUAGCCAGCCAACUCA |
| JA8 | GGCACGACUAGACAGCCAACUCA |
| JA9 | GGCACGAGACACCUACCGCCUCA |
| JA10 | GGCACGAGACAUCCAACGCCUCA |
| JA11 | GGCACGAGACACCCAACGACUCA |
| JA12 | GGCACGAGACAUGCCACACCGCACUCA |
| JA13 | GGCACGACUAGACACCUCACGACCUCA |
| JA14 | GGCACGACUAGACAUCUAACUCCCGUCA |
| JA15 | GGCACGAACUAGCCAGCAUUCGACUCGUCA |
| JA16 | GGCACGACUAUUCGACUCA |
| JA17 | GGCACGAACAUCGACGUCA |
| JA18 | GGCACGACUAGCCAGCUCA |
| JA19 | GGCACGACGAACGCCAUCA |
| JA20 | GGCACGACUAGACAGCCAUCA |
| JA21 | GGCACGAUAGCCAGCACUGCUCA |
| JA22 | GGCACGAACAGCCAUCCGACUCA |
| JA23 | GGCACGACUAGCCAGCCCCGUCA |

## Illumina Small RNA TruSeq Oligo sequences

| | |
|---|---|
| RNA 3' adapter RA3 #15013207 | 5-/5rApp/TGGAATTCTCGGGTGCCAAGG /3SpC3/-3 |
| RNA RT Primer (RTP), #15013981 | 5-GCCTTGGCACCCGAGAATTCCA-3 |
| **First round PCR Primer** | |
| Reverse polyT primer | 5-GTTCAGAGTTCTACAGTCCGACGATC**T** $_{(17)}$-3 |
| Reverse 5N polyT primer | 5-GTTCAGAGTTCTACAGTCCGACGATCNNNNN**T** $_{(20)}$-3 |
| Forward primer | 5-GCCTTGGCACCCGAGAATTCCA-3 |
| **Illumina sequencing primer (5'→3')** | |
| Universal primer | AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA |
| RPI1 | CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI2 | CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI3 | CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI4 | CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI5 | CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI6 | CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI7 | CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI8 | CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI9 | CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI10 | CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI11 | CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI12 | CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI13 | CAAGCAGAAGACGGCATACGAGAT**TTGACT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI14 | CAAGCAGAAGACGGCATACGAGAT**GGAACT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI15 | CAAGCAGAAGACGGCATACGAGAT**TGACAT**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |
| RPI16 | CAAGCAGAAGACGGCATACGAGAT**GGACGG**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA |

# References


Afonina I, Zivarts M, Kutyavin I, Lukhtanov E, Gamper H, Meyer RB. 1997. Efficient priming of PCR with short oligonucleotides conjugated to a minor groove binder. *Nucleic Acids Res* **25**: 2657–2660.

Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Gruissem W, Meins F, Hohn T, et al. 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Res* **34**: 462–471.

Alefelder S, Patel BK, Eckstein F. 1998. Incorporation of terminal phosphorothioates into oligonucleotides. *Nucleic Acids Res* **26**: 4983–8.

al-Karadaghi S, Aevarsson a, Garber M, Zheltonosova J, Liljas a. 1996. The structure of elongation factor G in complex with GDP: conformational flexibility and nucleotide exchange. *Structure* **4**: 555–565.

Anderson M, Schultz EP, Martick M, Scott WG. 2013. Active-Site Monovalent Cations Revealed in a 1.55-Å-Resolution Hammerhead Ribozyme Structure. *J Mol Biol* **425**: 3790–3798.

Baer MF, Reilly RM, McCorkle GM, Hai TY, Altman S, RajBhandary UL. 1988. The recognition by RNase P of precursor tRNAs. *J Biol Chem* **263**: 2344–51.

Bailey CH. 1938a. The Origin of Life (Oparin, A. I.). *J Chem Educ* **15**: 399.

Bailey CH. 1938b. The Origin of Life (Oparin, A. I.). *J Chem Educ* **15**: 399.

Balke D, Kuss A, Müller S. 2015. Landmarks in the Evolution of (t)-RNAs from the Origin of Life up to Their Present Role in Human Cognition. *Life (Basel, Switzerland)* **6**: 1.

Balme DM. 1962. Development of Biology in Aristotle and Theophrastus: Theory of Spontaneous Generation. *Phronesis* **7**: 91–104.

Banik SD, Nandi N. 2010. Aminoacylation reaction in the histidyl-tRNA synthetase: fidelity mechanism of the activation step. *J Phys Chem B* **114**: 2301–11.

Bao Y, Higgins L, Zhang P, Chan S, Laget S, Sweeney S, Lunnen K, Xu S. 2008. Expression and Purification of BmrI Restriction Endonuclease and Its N-terminal Cleavage Domain Variants. *Protein Expr Purif* **58**: 42–52.

Bartel DP, Szostak JW. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science (80- )* **261**: 1411–8.

Bashan a, Yonath a. 2005. Ribosome crystallography: catalysis and evolution of peptide-bond formation, nascent chain elongation and its co-translational folding. *Biochem Soc Trans* **33**: 488–92.

Basturea GN. 2013. Research Methods for Detection and Quantitation of RNA Modifications. *Mater Methods* **3**.

Behm-Ansmant I, Helm M, Motorin Y. 2011. Use of specific chemical reagents for detection of modified nucleotides in RNA. *J Nucleic Acids* **2011**: 408053.

Berg, Paul; FRED H. BERGMANN,E. J. OFENGAND AMD. 1961. Enzymic Synthesis of Amino Acyl of Ribonucleic Acid. *Synthesis (Stuttg)* **236**: 1748–1757.

Bernhardt HS, Tate WP. 2012. Primordial soup or vinaigrette: did the RNA world evolve at acidic pH? *Biol Direct* **7**: 4.

Bhaskaran H, Rodriguez-Hernandez a., Perona JJ. 2012. Kinetics of tRNA folding monitored by aminoacylation. *Rna* **18**: 569–580.

Biondi E, Poudyal RR, Forgy JC, Sawyer AW, Maxwell AWR, Burke DH. 2013. Lewis acid catalysis of phosphoryl transfer from a copper(II)-NTP complex in a kinase ribozyme. *Nucleic Acids Res* **41**:

3327–3338.
Boots JL, Canny MD, Azimi E, Pardi A. 2008. Metal ion specificities for folding and cleavage activity in the Schistosoma hammerhead ribozyme. *RNA* **14**: 2212–22.
Bowman JC, Hud N V., Williams LD. 2015. The Ribosome Challenge to the RNA World. *J Mol Evol* **80**: 143–161.
Bowman JC, Lenz TK, Hud N V., Williams LD. 2012. Cations in charge: Magnesium ions in RNA folding and catalysis. *Curr Opin Struct Biol* **22**: 262–272.
Breaker RR. 1997. In Vitro Selection of Catalytic Polynucleotides. *Chem Rev* **97**: 371–390.
Breaker RR, Joyce GF. 1994. Minimonsters: Evolutionary Byproducts of In Vitro RNA Amplification. In *Self-Production of Supramolecular Structures: From Synthetic Structures to Models of Minimal Living Systems* (eds. G.R. Fleischaker, S. Colonna, and P.L. Luisi), pp. 127–135, Springer Netherlands, Dordrecht.
Bruce AG, Uhlenbeck OC. 1978. Reactions at the termini of tRNA with T4 RNA ligase. *Nucleic Acids Res* **5**: 3665–3677.
Burke DH, Hoffman DC. 1998. A novel acidophilic RNA motif that recognizes coenzyme A. *Biochemistry* **37**: 4653–63.
Carothers JM, Davis JH, Chou JJ, Szostak JW. 2006. Solution structure of an informationally complex high-affinity RNA aptamer to GTP. *RNA* **12**: 567–79.
Carter C. 2015. What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life* **5**: 294–320.
Caruthers MH. 2013. The chemical synthesis of DNA/RNA: our gift to science. *J Biol Chem* **288**: 1420–1427.
Chan S, Bao Y, Ciszak E, Laget S, Xu S. 2007. Catalytic domain of restriction endonuclease BmrI as a cleavage module for engineering endonucleases with novel substrate specificities. *Nucleic Acids Res* **35**: 6238–6248.
Chen D, Patton JT. 2001. Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. *Biotechniques* **30**: 574–582.
Chen IA, Salehi-Ashtiani K, Szostak JW. 2005. RNA Catalysis in Model Protocell Vesicles. *J Am Chem Soc* **127**: 13213–13219.
Chen X, Li N, Ellington AD. 2007. Ribozyme catalysis of metabolism in the RNA world. *Chem Biodivers* **4**: 633–55.
Chen X, Sim S, Wurtmann EJ, Feke A, Wolin SL. 2014. Bacterial noncoding Y RNAs are widespread and mimic tRNAs. *RNA* **20**: 1715–24.
Chien A, Edgar DB, Trela JM. 1976. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J Bacteriol* **127**: 1550–1557.
Chumachenko N V, Novikov Y, Yarus M. 2009. Rapid and Simple Ribozymic Aminoacylation Using Three Conserved Nucleotides. *J Am Chem Soc* **131**: 5257–5263.
Connell GJ, Christian EL. 1993. Utilization of cofactors expands metabolism in a new RNA world. *Orig Life Evol Biosph* **23**: 291–297.
Copley SD, Smith E, Morowitz HJ. 2007. The origin of the RNA world: Co-evolution of genes and metabolism. *Bioorg Chem* **35**: 430–443.
Corfù N a, Sigel H. 1991. Acid-base properties of nucleosides and nucleotides as a function of concentration. Comparison of the proton affinity of the nucleic base residues in the monomeric and self-associated, oligomeric 5'-triphosphates of inosine (ITP), guanosine (GTP), and ade. *Eur J Biochem* **199**: 659–669.

Cotten M, Birnstiel ML. 1989. Ribozyme mediated destruction of RNA in vivo. *EMBO J* **8**: 3861–3866.
Curtis EA, Bartel DP. 2005. New catalytic structures from an existing ribozyme. *Nat Struct Mol Biol* **12**: 994–1000.
da Silva J a. L. 2015. From the RNA world to the RNA/protein world: Contribution of some riboswitch-binding species? *J Theor Biol* **370**: 197–201.
Dange V, Van Atta RB, Hecht SM. 1990. A Mn2(+)-dependent ribozyme. *Science* **248**: 585–8.
DeRose VJ. 2003. Metal ion binding to catalytic RNA molecules. *Curr Opin Struct Biol* **13**: 317–324.
DeRose VJ. 2002. Two decades of RNA catalysis. *Chem Biol* **9**: 961–969.
Desogus G, Todone F, Brick P, Onesti S. 2000. Active site of lysyl-tRNA synthetase: structural studies of the adenylation reaction. *Biochemistry* **39**: 8418–25.
Diebel KW, Zhou K, Clarke AB, Bemis LT. 2016. Beyond the Ribosome: Extra-translational Functions of tRNA Fragments. *Biomark Insights* **11**: 1–8.
Dieckmann T, Suzuki E, Nakamura GK, Feigon J. 1996. Solution structure of an ATP-binding RNA aptamer reveals a novel fold. *RNA* **2**: 628–640.
Dittmar K a, Sørensen M a, Elf J, Ehrenberg M, Pan T. 2005. Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep* **6**: 151–7.
Dolan GF, Akoopie A, Müller UF. 2015. A Faster Triphosphorylation Ribozyme ed. A.S. Lewin. *PLoS One* **10**: e0142559.
Dong H, Nilsson L, Kurland CG. 1996a. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**: 649–63.
Dong H, Nilsson L, Kurland CG. 1996b. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**: 649–63.
Dreher TW. 2009. Role of tRNA-like structures in controlling plant virus replication. *Virus Res* **139**: 217–29.
Dupont DM, Larsen N, Jensen JK, Andreasen PA, Kjems J. 2015. Characterisation of aptamer–target interactions by branched selection and high-throughput sequencing of SELEX pools. *Nucleic Acids Res* **43**: gkv700.
Eigen M, Schuster P. 1978. The Hypercycle. *Naturwissenschaften* **65**: 7–41.
Ekland EH, Bartel DP. 1996. RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* **382**: 373–376.
Ekland EH, Szostak JW, Bartel DP. 1995. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**: 364–70.
El Yacoubi B, Bailly M, de Crécy-Lagard V. 2012. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu Rev Genet* **46**: 69–95.
Ellington AD, Chen X, Robertson M, Syrett A. 2009. Evolutionary origins and directed evolution of RNA. *Int J Biochem Cell Biol* **41**: 254–265.
England TE, Uhlenbeck OC. 1978. Enzymatic oligoribonucleotide synthesis with T4 RNA ligase. *Biochemistry* **17**: 2069–76.
Erat MC, Sigel RKO. 2008. Divalent metal ions tune the self-splicing reaction of the yeast mitochondrial group II intron Sc.ai5?? *J Biol Inorg Chem* **13**: 1025–1036.
Ezraty B, Dahlgren B, Deutscher MP. 2005. The RNase Z Homologue Encoded by Escherichia coli elaC Gene Is RNase BN. *J Biol Chem* **280**: 16542–16545.
Famulok M. 1994. Molecular Recognition of Amino Acids by RNA-Aptamers: An L-Citrulline Binding RNA Motif and Its Evolution into an L-Arginine Binder. *J Am Chem Soc* **116**: 1698–1706.
Fechter P, Rudinger-Thirion J, Florentz C, Giegé R. 2001. Novel features in the tRNA-like world of plant

viral RNAs. *Cell Mol Life Sci* **58**: 1547–1561.
Fiorini E, Börner R, Sigel RKO. 2015. Mimicking the in vivo Environment--The Effect of Crowding on RNA and Biomacromolecular Folding and Activity. *Chimia (Aarau)* **69**: 207–12.
Flores R, Gago-Zachert S, Serra P, Sanjuán R, Elena SF. 2014. Viroids: Survivors from the RNA World? *Annu Rev Microbiol* 395–414.
Francklyn C, Schimmel P. 1990. Enzymatic aminoacylation of an eight-base-pair microhelix with histidine. *Proc Natl Acad Sci U S A* **87**: 8655–9.
Fraser LA, Kinghorn AB, Tang MSL, Cheung YW, Lim B, Liang S, Dirkzwager RM, Tanner JA, Miller AOA, Vanden Eynde JJ. 2015. Oligonucleotide functionalised microbeads: Indispensable tools for high-throughput aptamer selection. *Molecules* **20**: 21298–21312.
Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A* **85**: 8998–9002.
Fu Y, Lee I, Lee YS, Bao X. 2015. Small Non-coding Transfer RNA-Derived RNA Fragments (tRFs): Their Biogenesis, Function and Implication in Human Diseases. *Genomics Inform* **13**: 94–101.
Fujita Y, Furuta H, Ikawa Y. 2010. Evolutionary optimization of a modular ligase ribozyme: a small catalytic unit and a hairpin motif masking an element that could form an inactive structure. *Nucleic Acids Res* **38**: 3328–39.
Garcia-Silva MR, Cabrera-Cabrera F, G??ida MC, Cayota A. 2012. Hints of tRNA-derived small RNAs role in RNA silencing mechanisms. *Genes (Basel)* **3**: 603–614.
Gaston KW, Rubio MAT, Alfonzo JD. 2008. OXOPAP assay: for selective amplification of aminoacylated tRNAs from total cellular fractions. *Methods* **44**: 170–5.
Geiger A, Burgstaller P, Von der Eltz H, Roeder A, Famulok M. 1996. RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res* **24**: 1029–1036.
Gellert M, Lipsett MN, Davies DR. 1962. HELIX FORMATION BY GUANYLIC ACID. *Proc Natl Acad Sci U S A* **48**: 2013–2018.
Geslain R, De Pouplana LR. 2004. Regulation of RNA function by aminoacylation and editing? *Trends Genet* **20**: 604–610.
Giegé R, Springer M. 2012. Aminoacyl-tRNA Synthetases in the Bacterial World. *EcoSal Plus* **5**.
Gilbert W. 1986. The RNA world. *Nature* **319**: 618.
Green NJ, Grundy FJ, Henkin TM. 2010. The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett* **584**: 318–324.
GROSSENBACHER KA, KNIGHT CA. 1965. AMINO ACIDS, PEPTIDES, AND SPHERULES OBTAINED FROM "PRIMITIVE EARTH" GASES IN A SPARKING SYSTEM A2 - FOX, SIDNEY W. BT - The Origins of Prebiological Systems and of their Molecular Matrices. pp. 173–186, Academic Press.
Guo Y, Bosompem A, Mohan S, Erdogan B, Ye F, Vickers KC, Sheng Q, Zhao S, Li C-I, Su P-F, et al. 2015. Transfer RNA detection by small RNA deep sequencing and disease association with myelodysplastic syndromes. *BMC Genomics* **16**: 727.
Hager AJ, Szostak JW. 1997. Isolation of novel ribozymes that ligate AMP-activated RNA substrates. *Chem Biol* **4**: 607–617.
Hanna R, Doudna JA. 2000. Metal ions in ribozyme folding and catalysis. *Curr Opin Chem Biol* **4**: 166–170.
Harrison B, Zimmerman SB. 1984. Polymer-stimulated ligation: enhanced ligation of oligo- and

polynucleotides by T4 RNA ligase in polymer solutions. *Nucleic Acids Res* **12**: 8235–8251.
Harvey RJ, Darlison MG. 1991. Random-primed cDNA synthesis facilitates the isolation of multiple 5'-cDNA ends by RACE. *Nucleic Acids Res* **19**: 4002.
Hatano M, Ishihara K. 2013. Lanthanum(III) catalysts for highly efficient and chemoselective transesterification. *Chem Commun (Camb)* **49**: 1983–97.
Hema M, Gopinath K, Kao C. 2005. Repair of the tRNA-Like CCA Sequence in a Multipartite Positive-Strand RNA Virus. *J Virol* **79**: 1417–1427.
Henkin TM. 2008a. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22**: 3383–90.
Henkin TM. 2008b. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22**: 3383–90.
Hernandez AR, Piccirilli JA. 2013. Chemical origins of life: Prebiotic RNA unstuck. *Nat Chem* **5**: 360–362.
Higgs PG, Lehman N. 2014. The RNA World: molecular cooperation at the origins of life. *Nat Rev Genet* **16**: 7–17.
Holm RH, Kennepohl P, Solomon EI. 1996. Structural and Functional Aspects of Metal Sites in Biology. *Chem Rev* **96**: 2239–2314.
Huang F, Yarus M. 1997. 5'-RNA self-capping from guanosine diphosphate. *Biochemistry* **36**: 6557–63.
Huang Z, Szostak JW. 2003. Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. *RNA* **9**: 1456–1463.
Hughes G, Nevell T. 1948. the mechanism of the oxidation of glucose by periodate. *Trans Faraday Soc*.
Hyndman DL, Mitsuhashi M. 2003. PCR Protocols. In (eds. J.M.S. Bartlett and D. Stirling), pp. 81–88, Humana Press, Totowa, NJ.
Ibba M. 2015. Transfer RNA comes of age. *Rna* 648–649.
Illangasekare M, Sanchez G, Nickles T, Yarus M. 1995. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science (80- )*.
Illangasekare M, Yarus M. 1997. Small-molecule-Substrate Interactions with a Self-aminoacylating ribozyme. *J Mol Biol* 631–639.
Illangasekare M, Yarus M. 1999. Specific, rapid synthesis of Phe-RNA by RNA. *Proc Natl Acad Sci U S A* **96**: 5470–5.
Ivery TC, Daron HH, Aull JL. 1984. The inactivation of thymidylate synthase by periodate. *J Inorg Biochem* **22**: 259–270.
Janssen BD, Hayes CS. 2012. The tmRNA ribosome-rescue system. *Adv Protein Chem Struct Biol* **86**: 151–91.
Jäschke A, Seelig B. 2000. Evolution of DNA and RNA as catalysts for chemical reactions. *Curr Opin Chem Biol* 257–262.
Jayasena VK, Gold L. 1997. In vitro selection of self-cleaving RNAs with a low pH optimum. *Proc Natl Acad Sci U S A* **94**: 10612–7.
Jenne a, Famulok M. 1998. A novel ribozyme with ester transferase activity. *Chem Biol* **5**: 23–34.
Jiang Y-F, Xiao M, Yin P, Zhang Y. 2006. Monovalent cations use multiple mechanisms to resolve ribozyme misfolding. *RNA* **12**: 561–566.
Jimenez RM, Polanco JA, Lupták A. 2015. Chemistry and Biology of Self-Cleaving Ribozymes. *Trends Biochem Sci* **40**: 648–661.
Johnson-Buck AE, McDowell SE, Walter NG. 2011. Metal ions: supporting actors in the playbook of small ribozymes. *Met Ions Life Sci* **9**: 175–96.
Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**: 1319–25.

Joyce GF. 1994. In vitro evolution of nucleic acids. *Curr Opin Struct Biol* **4**: 331–336.
Joyce GF. 1989. RNA evolution and the origins of life. *Nature* **338**: 217–224.
Joyce GF. 2002. The antiquity of RNA-based evolution. *Nature* **418**: 214–221.
Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J. 2009. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* **37**: D159–D162.
Kanai A. 2015. Disrupted tRNA Genes and tRNA Fragments: A Perspective on tRNA Gene Evolution. *Life* **5**: 321–331.
Karimata H, Nakano S, Sugimoto N. 2006. The roles of cosolutes on the hammerhead ribozyme activity. *Nucleic Acids Symp Ser (Oxf)* 81–2.
KAUFMANN G, KALLENBACH NR. 1975. Determination of recognition sites of T4 RNA ligase on the 3[prime]-OH and 5[prime]-P termini of polyribonucleotide chains. *Nature* **254**: 452–454.
Kawano M, Kawaji H, Grandjean V, Kiani J, Rassoulzadegan M. 2012. Novel small noncoding RNAs in mouse spermatozoa, zygotes and early embryos. *PLoS One* **7**: e44542.
Ke A, Ding F, Batchelor JD, Doudna JA. 2007. Structural Roles of Monovalent Cations in the HDV Ribozyme. *Structure* **15**: 281–287.
Keam S, Hutvagner G. 2015. tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life* **5**: 1638–1651.
Kilburn D, Roh JH, Behrouzi R, Briber RM, Woodson S a. 2013. Crowders perturb the entropy of RNA energy landscapes to favor folding. *J Am Chem Soc* **135**: 10055–63.
Kisseleva N, Khvorova A, Westhof E, Schiemann O. 2005. Binding of manganese ( II ) to a tertiary stabilized hammerhead ribozyme as studied by electron paramagnetic resonance spectroscopy. 1–6.
Koizumi M, Breaker RR. 2000. Molecular recognition of cAMP by an RNA aptamer. *Biochemistry* **39**: 8983–8992.
Kolev NG, Hartland EI, Huber PW. 2008. A manganese-dependent ribozyme in the 3'-untranslated region of Xenopus Vg1 mRNA. *Nucleic Acids Res* **36**: 5530–5539.
Komine Y, Kitabatake M, Yokogawa T, Nishikawa K, Inokuchi H. 1994. A tRNA-like structure is present in 10Sa RNA, a small stable RNA from Escherichia coli. *Proc Natl Acad Sci U S A* **91**: 9223–7.
Könnyű B, Szilágyi A, Czárán T. 2015. In silico ribozyme evolution in a metabolically coupled RNA population. *Biol Direct* **10**: 30.
Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**: 147–157.
Krupp G. 1988. RNA synthesis: strategies for the use of bacteriophage RNA polymerases. *Gene* **72**: 75–89.
Kulpa D, Topping R, Telesnitsky A. 1997. Determination of the site of rst strand transfer during Moloney murine leukemia virus reverse transcription and identi cation of strand transfer-associated reverse transcriptase errors. *Cancer* **16**: 856–865.
Kumar PK, Ellington AD. 1995. Artificial evolution and natural ribozymes. *FASEB J* **9**: 1183–95.
Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, Maloof JN, Sinha NR. 2012. A High-Throughput Method for Illumina RNA-Seq Library Preparation. *Front Plant Sci* **3**: 1–10.
Kumar RK, Yarus M. 2001. RNA-catalyzed amino acid activation. *Biochemistry* **40**: 6998–7004.
Kurata S. 2003. Quick two-step RNA ligation employing periodate oxidation. *Nucleic Acids Res* **31**: 145e–145.
Lancaster AM, Jan E, Sarnow P. 2006. Initiation factor-independent translation mediated by the hepatitis C virus internal ribosome entry site. *RNA* **12**: 894–902.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short

DNA sequences to the human genome. *Genome Biol* **10**: R25.
Lazinski DW, Camilli A. 2013. Homopolymer tail-mediated ligation PCR: A streamlined and highly efficient method for DNA cloning and library construction. *Biotechniques* **54**: 25–34.
Lee DH, Granja JR, Martinez JA, Severin K, Ghadiri MR. 1996. A self-replicating peptide. *Nature* **382**: 525–528.
Lee H-T, Kilburn D, Behrouzi R, Briber RM, Woodson SA. 2015. Molecular crowding overcomes the destabilizing effects of mutations in a bacterial ribozyme. *Nucleic Acids Res* **43**: 1170–6.
Lee N, Bessho Y, Wei K, Szostak JW, Suga H. 2000. Ribozyme-catalyzed tRNA aminoacylation. *Nat Struct Biol* **7**: 28–33.
Lee TS, Wong KY, Giambasu GM, York DM. 2013. *Bridging the gap between theory and experiment to derive a detailed understanding of hammerhead ribozyme catalysis*. 1st ed. Elsevier Inc.
Lee YS, Shibata Y, Malhotra A, Lee YS, Shibata Y, Malhotra A, Dutta A. 2009. A novel class of small RNAs : tRNA-derived RNA fragments ( tRFs ) A novel class of small RNAs : tRNA-derived RNA fragments ( tRFs ). 2639–2649.
Lehmann J, Reichel A, Buguin A, Libchaber A. 2007. Efficiency of a self-aminoacylating ribozyme: effect of the length and base-composition of its 3' extension. *RNA* **13**: 1191–7.
Li N, Huang F. 2005. Ribozyme-catalyzed aminoacylation from CoA thioesters. *Biochemistry* **44**: 4582–90.
Lie L, Biliya S, Vannberg F, Wartell RM. 2016. Ligation of RNA Oligomers by the Schistosoma mansoni Hammerhead Ribozyme in Frozen Solution. *J Mol Evol* **82**: 81–92.
Lightfoot DA. 1988. Magnesium-dependence of in vitro translation programmed by gene-specific mRNAs. *Nucleic Acids Res* **16**: 4164.
Linzell JL, Peaker M. 1971. Intracellular concentrations of sodium, potassium and chloride in the lactating mammary gland and their relation to the secretory mechanism. *J Physiol* **216**: 683–700.
Lohse P, Szostak J. 1996. Ribozyme-catalysed amino-acid transfer reactions. *Nature*.
Lönnberg T, Lönnberg H. 2005. Chemical models for ribozyme action. *Curr Opin Chem Biol* **9**: 665–673.
Loring H, Levy L. 1956. Periodate Oxidation of Sugar Phosphates in Neutral Solution. I. D-Ribose 5-Phosphate1. *J Am …* **78**: 3724.
Lorsch JR, Szostak JW. 1994. In vitro evolution of new ribozymes with polynucleotide kinase activity. *Nature* **371**: 31–6.
M.A. Innis, D.H. Gelfand JJS and TJW. 1990. *PCR protocols — A guide to methods and applications*. Academic Press.
M. Illangasekare. 1999. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA* **5**: 1482–1489.
Majerfeld I, Yarus M. 2005. A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res* **33**: 5482–5493.
Mannironi C, Scerch C, Fruscoloni P, Tocchini-Valentini GP. 2000. Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif. *RNA* **6**: 520–527.
Marinetti G, Rouser G. 1955. The periodate oxidation of ribose-5-phosphate in acid and alkaline solution. *… Am Chem Soc*.
Martin L, Unrau P, Müller U. 2015. RNA Synthesis by in Vitro Selected Ribozymes for Recreating an RNA World. *Life* **5**: 247–268.
McCutchan TF, Hansen JL, Dame JB, Mullins JA. 1984. Mung bean nuclease cleaves Plasmodium genomic DNA at sites before and after genes. *Science (80- )* **225**: 625–628.

MILLER SL. 1953. A production of amino acids under possible primitive earth conditions. *Science* **117**: 528–9.

Milligan JF, Groebe DR, Witherell GW, Uhlenbeck OC. 1987. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res* **15**: 8783–8798.

Mills DR, Peterson RL, Spiegelman S. 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc Natl Acad Sci U S A* **58**: 217–224.

Miyamoto Y, Teramoto N, Imanishi Y, Ito Y. 2005. In vitro evolution and characterization of a ligase ribozyme adapted to acidic conditions: effect of further rounds of evolution. *Biotechnol Bioeng* **90**: 36–45.

Moore MJ, Sharp PA. 1992. Site-specific modification of pre-mRNA: the 2'-hydroxyl groups at the splice sites. *Science (80- )* **256**: 992–997.

Moretti JE, Müller UF. 2014. A ribozyme that triphosphorylates RNA 5'-hydroxyl groups. *Nucleic Acids Res* 1–12.

Morii T, Hagihara M, Sato S, Makino K. 2002. In vitro selection of ATP-binding receptors using a ribonucleopeptide complex. *J Am Chem Soc* **124**: 4617–22.

Morimoto J, Hayashi Y, Iwasaki K, Suga H. 2011. Flexizymes: their evolutionary history and the origin of catalytic function. *Acc Chem Res* **44**: 1359–68.

Morowitz HJ, Heinz B, Deamer DW. 1988a. The chemical logic of a minimum protocell. *Orig life Evol Biosph* **18**: 281–287.

Morowitz HJ, Heinz B, Deamer DW. 1988b. The chemical logic of a minimum protocell. *Orig life Evol Biosph* **18**: 281–287.

Motea EA, Berdis AJ. 2010. Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. *Biochim Biophys Acta - Proteins Proteomics* **1804**: 1151–1166.

Motorin Y, Helm M. 2011. RNA nucleotide methylation. *Wiley Interdiscip Rev RNA* **2**: 611–631.

Motorin Y, Muller S, Behm-Ansmant I, Branlant C. 2007. Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol* **425**: 21–53.

Munafó DB, Robb GB. 2010. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA* **16**: 2537–2552.

Murakami H, Bonzagni NJ, Suga H. 2002. Aminoacyl-tRNA synthesis by a resin-immobilized ribozyme. *J Am Chem Soc* **124**: 6834–5.

Murray JB, Dunham CM, Scott WG. 2002. A pH-dependent conformational change, rather than the chemical step, appears to be rate-limiting in the hammerhead ribozyme cleavage reaction. *J Mol Biol* **315**: 121–130.

Murray JB, Seyhan a a, Walter NG, Burke JM, Scott WG. 1998. The hammerhead, hairpin and VS ribozymes are catalytically proficient in monovalent cations alone. *Chem Biol* **5**: 587–595.

Murtas G. 2013. Early self-reproduction, the emergence of division mechanisms in protocells. *Mol Biosyst* **9**: 195–204.

Nakano S, Karimata HT, Kitagawa Y, Sugimoto N. 2009. Facilitation of RNA enzyme activity in the molecular crowding media of cosolutes. *J Am Chem Soc* **131**: 16881–8.

Nakano S, Kitagawa Y, Karimata HT, Sugimoto N. 2008. Molecular crowding effect on metal ion binding properties of the hammerhead ribozyme. *Nucleic Acids Symp Ser (Oxf)* 519–20.

Ninomiya K, Minohata T, Nishimura M, Sisido M. 2004. In situ chemical aminoacylation with amino acid thioesters linked to a peptide nucleic acid. *J Am Chem Soc* **126**: 15984–9.

Niwa N, Yamagishi Y, Murakami H, Suga H. 2009. A flexizyme that selectively charges amino acids activated by a water-friendly leaving group. *Bioorg Med Chem Lett* **19**: 3892–4.

Ohkawa J, Koguma T, Kohda T, Taira K. 1995. Ribozymes: from mechanistic studies to applications in vivo. *J Biochem* **118**: 251–8.

Ohtsuka E, Nishikawa S. 1976. Joining of ribooligonucleotides with T4 RNA ligase and identification of the oligonucleotide-adenylate intermediate. *Nucleic acids …* **3**: 1613–1623.

Ohuchi SP, Ikawa Y, Nakamura Y. 2008. Selection of a novel class of RNA-RNA interaction motifs based on the ligase ribozyme with defined modular architecture. *Nucleic Acids Res* **36**: 3600–7.

Orgel LE. 1986. RNA catalysis and the origins of life. *J Theor Biol* **123**: 127–149.

Orgel LE. 2003. Some consequences of the RNA world hypothesis. *Orig Life Evol Biosph* **33**: 211–8.

ORO J. 2002. Historical Understanding of Life's Beginnings. In *Life's Origin, The Beginnings of Biological Evolution*, pp. 7–45, University of California Press.

Oz-Gleenberg I, Herschhorn A, Hizi A. 2011. Reverse transcriptases can clamp together nucleic acids strands with two complementary bases at their 3???-termini for initiating DNA synthesis. *Nucleic Acids Res* **39**: 1042–1053.

Oz-Gleenberg I, Herzig E, Hizi A. 2012. Template-independent DNA synthesis activity associated with the reverse transcriptase of the long terminal repeat retrotransposon Tf1. *FEBS J* **279**: 142–53.

Pan T, Uhlenbeck OC. 1992. In vitro selection of RNAs that undergo autolytic cleavage with lead(2+). *Biochemistry* **31**: 3887–3895.

Pascal R, Boiteau L, Commeyras A. 2005. From the Prebiotic Synthesis of α-Amino Acids Towards a Primitive Translation Apparatus for the Synthesis of Peptides. In *Prebiotic Chemistry* (ed. P. Walde), pp. 69–122, Springer Berlin Heidelberg, Berlin, Heidelberg.

Perrotta AT, Been MD. 2007. A single nucleotide linked to a switch in metal ion reactivity preference in the HDV ribozymes. *Biochemistry* **46**: 5124–5130.

Perrotta AT, Been MD. 2006. HDV ribozyme activity in monovalent cations. *Biochemistry* **45**: 11357–11365.

Petrov AS, Bernier CR, Hsiao C, Okafor CD, Tannenbaum E, Stern J, Gaucher E, Schneider D, Hud N V, Harvey SC, et al. 2012. RNA-magnesium-protein interactions in large ribosomal subunit. *J Phys Chem B* **116**: 8113–20.

Pobanz K, Lupták A. 2016. Improving the odds: Influence of starting pools on in vitro selection outcomes. *Methods*.

Polisson C, Morgan RD. 1988. EarI, a restriction endonuclease from Enterobacter aerogenes which recognizes 5'CTCTTC3'. *Nucleic Acids Res* **16**: 9872.

Pontius BW, Lott WB, von Hippel PH. 1997. Observations on catalysis by hammerhead ribozymes are consistent with a two-divalent-metal-ion mechanism. *Proc Natl Acad Sci U S A* **94**: 2290–2294.

Pressman A, Blanco C, Chen IA. 2015. The RNA World as a Model System to Study the Origin of Life. *Curr Biol* **25**: R953–R963.

Puerto-Galán L, Vioque A. 2012. Expression and processing of an unusual tRNA gene cluster in the cyanobacterium Anabaena sp. PCC 7120. *FEMS Microbiol Lett* **337**: 10–7.

Quail MA, Swerdlow H, Turner DJ. 2009. Improved Protocols for the Illumina Genome Analyzer Sequencing System. In *Current Protocols in Human Genetics*, Vol. 18 of, p. 2, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Raghavan R, Hicks LD, Minnick MF. 2009. A Unique Group I Intron in Coxiella burnetii Is a Natural Splice Mutant . *J Bacteriol* **191**: 4044–4046.

Raina M, Ibba M. 2014. TRNAs as regulators of biological processes. *Front Genet* **5**: 1–14.

Ralser M. 2014. The RNA world and the origin of metabolic enzymes. *Biochem Soc Trans* **42**: 985–8.

Retailleau P, Weinreb V, Hu M, Carter CW. 2007. Crystal Structure of Tryptophanyl-tRNA Synthetase

Complexed with Adenosine-5′ Tetraphosphate: Evidence for Distributed Use of Catalytic Binding Energy in Amino Acid Activation by Class I Aminoacyl-tRNA Synthetases. *J Mol Biol* **369**: 108–128.

Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**: 8627–8637.

Riccitelli N, Lupták A. 2013. *HDV family of self-cleaving ribozymes*.

Robertson MP, Joyce GF, Noller HF, Volpe T, Martienssen RA. 2012. The Origins of the RNA World. *Cold Spring Harb Perspect Biol*.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Romaniuk E, McLaughlin LW, Neilson T, Romaniuk PJ. 1982. The effect of acceptor oligoribonucleotide sequence on the T4 RNA ligase reaction. *Eur J Biochem* **125**: 639–43.

Ross W, Vrentas CE, Sanchez-Vazquez P, Gaal T, Gourse RL. 2013. The Magic Spot: A ppGpp Binding Site on E. coli RNA Polymerase Responsible for Regulation of Transcription Initiation. *Mol Cell* **50**: 420–429.

Roth a, Breaker RR. 1998. An amino acid as a cofactor for a catalytic polynucleotide. *Proc Natl Acad Sci U S A* **95**: 6027–31.

Roychowdhury-Saha M, Burke DH. 2006. Extraordinary rates of transition metal ion-mediated ribozyme catalysis. *RNA* **12**: 1846–1852.

Ruckman J, Green LS, Beeson J, Waugh S, Gillette WL, Henninger DD, Claesson-Welsh L, Janjic N. 1998. 2′-Fluoropyrimidine RNA-based Aptamers to the 165-Amino Acid Form of Vascular Endothelial Growth Factor (VEGF165): INHIBITION OF RECEPTOR BINDING AND VEGF-INDUCED VASCULAR PERMEABILITY THROUGH INTERACTIONS REQUIRING THE EXON 7-ENCODED DOMAIN . *J Biol Chem* **273**: 20556–20567.

Ryu KH, Choi SH, Lee JS. 2000. Restriction primers as short as 6-mers for PCR amplification of bacterial and plant genomic DNA and plant viral RNA. *Mol Biotechnol* **14**: 1–3.

Saito H, Kourouklis D, Suga H. 2001. An in vitro evolved precursor tRNA with aminoacylation activity. *EMBO J* **20**: 1797–806.

Saito H, Suga H. 2002. Outersphere and innersphere coordinated metal ions in an aminoacyl-tRNA synthetase ribozyme. *Nucleic Acids Res* **30**: 5151–5159.

Saran D, Nickens DG, Burke DH. 2005. A trans acting ribozyme that phosphorylates exogenous RNA. *Biochemistry* **44**: 15007–15016.

Sassanfar M, Szostak JW. 1993. An RNA motif that binds ATP. *Nature* **364**: 550–553.

Sazani PL, Larralde R, Szostak JW. 2004. A small aptamer with strong and specific recognition of the triphosphate of ATP. *J Am Chem Soc* **126**: 8370–1.

Schimmel P, Kelley SO. 2000. Exiting an RNA world. *Nat Struct Biol* **7**: 5–7.

Schnabl J, Sigel RKO. 2010. Controlling ribozyme activity by metal ions. *Curr Opin Chem Biol* **14**: 269–75.

Schürer H, Lang K, Schuster J, Mörl M. 2002. A universal method to produce in vitro transcripts with homogeneous 3' ends. *Nucleic Acids Res* **30**: e56.

Schutz K, Hesselberth JR, Fields S. 2010. Capture and sequence analysis of RNAs with terminal 2′,3′-cyclic phosphates. *RNA* **16**: 621–631.

Schwartz W. 1971. Melvin Calvin, Chemical Evolution. Molecular Evolution towards the Origin of Living Systems on the Earth and elsewhere. IX und 278 S., 154 Abb., 24 Tab., 8 Taf. Oxford 1969: Clarendon Press 55 s. *Z Allg Mikrobiol* **11**: 256.

Scott WG, Klug A. 1996. Ribozymes: Structure and mechanism in RNA catalysis. *Trends Biochem Sci* **21**: 220–224.

Shapiro R. 1987. *Origins: A skeptic's guide to the creation of life on earth*. Bantam Dell Pub Group.

Shi J, Martinis S, Schimmel P. 1992. RNA tetraloops as minimalist substrates for aminoacylation. *Biochemistry* 4931–4936.

Sigel RKO, Pyle AM. 2007. Alternative roles for metal ions in enzyme catalysis and the implications for ribozyme chemistry. *Chem Rev* **107**: 97–113.

Sleeper H, Orgel L. 1979. The catalysis of nucleotide polymerization by compounds of divalent lead. *J Mol Evol* **12**: 357–364.

Sninsky JJ, Last JA, Gilham PT. 1976. The use of terminal blocking groups for the specific joining of oligonucleotides in RNA ligase reactions containing equimolar concentrations of acceptor and donor molecules. *Nucleic Acids Res* **3**: 3157–3166.

Soutourina J, Plateau P, Delort F, Peirotes A, Blanquet S. 1999. Functional Characterization of the D -Tyr-tRNA Tyr Deacylase from Escherichia coli *. *Biochemistry* **274**: 19109–19114.

Stolze K, Holmes SC, Earnshaw DJ, Singh M, Stetsenko D, Williams D, Gait MJ. 2001. Novel spermine-amino acid conjugates and basic tripeptides enhance cleavage of the hairpin ribozyme at low magnesium ion concentration. *Bioorg Med Chem Lett* **11**: 3007–10.

Sufrin J, Spiess A, Jr CM. 1995. Purine 2′, 3′-acyclonucleosides: Improved synthesis and antiparasitic actimity. *Bioorganic Med …* **5**: 1961–1964.

Sugino A, Snoper TJ, Cozzarelli NR. 1977. Bacteriophage T4 RNA ligase. Reaction intermediates and interaction of substrates. *J Biol Chem* **252**: 1732–1738.

Sun L, Cui Z, Gottlieb RL, Zhang B. 2002. A selected ribozyme catalyzing diverse dipeptide synthesis. *Chem Biol* **9**: 619–28.

Sun T, Zhang Y. 2008. Pentamidine binds to tRNA through non-specific hydrophobic interactions and inhibits aminoacylation and translation. *Nucleic Acids Res* **36**: 1654–1664.

Takaku H, Nashimoto M. 2008. Escherichia coli tRNase Z can shut down growth probably by removing amino acids from aminoacyl-tRNAs. *Genes to Cells*.

Tang J, Breaker RR. 1998. Mechanism for allosteric inhibition of an ATP-sensitive ribozyme. *Nucleic Acids Res* **26**: 4214–4221.

Tateishi-Karimata H, Muraoka T, Kinbara K, Sugimoto N. 2016. G-quadruplexes with tetraethylene glycol-modified deoxythymidines are resistant to nucleases and inhibit HIV-1 reverse transcriptase. *ChemBioChem* n/a--n/a.

Teramoto N, Imanishi Y, Ito Y. 2000. In Vitro Selection of a Ligase Ribozyme Carrying Alkylamino Groups in the Side Chains. *Bioconjug Chem* **11**: 744–748.

Thiel WH, Bair T, Wyatt Thiel K, Dassie JP, Rockey WM, Howell CA, Liu XY, Dupuy AJ, Huang L, Owczarzy R, et al. 2011. Nucleotide bias observed with a short SELEX RNA aptamer library. *Nucleic Acid Ther* **21**: 253–63.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

Tjivikua T, Ballester P, Rebek J. 1990. Self-replicating system. *J Am Chem Soc* **112**: 1249–1250.

Turk RM, Chumachenko N V, Yarus M. 2010. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci U S A* **107**: 4585–9.

Turk RM, Illangasekare M, Yarus M. 2011. Catalyzed and spontaneous reactions on ribozyme ribose. *J Am Chem Soc* **133**: 6044–50.

Turunen JJ, Pavlova L V, Hengesbach M, Helm M, Müller S, Hartmann RK, Frilander MJ. 2014. RNA

Ligation. In *Handbook of RNA Biochemistry*, pp. 45–88, Wiley-VCH Verlag GmbH & Co. KGaA.
Ucisik MN, Bevilacqua PC, Hammes-Schiffer S. 2016. Molecular Dynamics Study of Twister Ribozyme: Role of Mg2+ Ions and Hydrogen-Bonding Network in Active Site. *Biochemistry*.
Uhlenbeck OC, Cameron V. 1977. Equimolar addition of oligoribonucleotides with T4 RNA ligase. *Nucleic Acids Res* **4**: 85–98.
Uzman A. 2003. Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Biochem Mol Biol Educ* **31**: 212–214.
Uzman A. 2001. Molecular Cell Biology (4th edition) Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore and James Darnell; Freeman & Co., New York, NY, 2000, 1084 pp., list price $102.25, ISBN 0-7167-3136-3. *Biochem Mol Biol Educ* **29**: 126–128.
Vaish NK, Larralde R, Fraley AW, Szostak JW, McLaughlin LW. 2003. A novel, modification-dependent ATP-binding aptamer selected from an RNA library incorporating a cationic functionality. *Biochemistry* **42**: 8842–51.
van der Gulik P, Speijer D. 2015. How Amino Acids and Peptides Shaped the RNA World. *Life* **5**: 230–246.
Vincent J, Gurling H, Melmer G. 1991. Oligonucleotides As Short As 7-Mers Can Be Used for Pcr Amplification. **13**: 75–82.
Vourekas A, Stamatopoulou V, Toumpeki C, Tsitlaidou M, Drainas D. 2008. Insights into functional modulation of catalytic RNA activity. *IUBMB Life* **60**: 669–683.
Westheimer F. 1987. Why nature chose phosphates. *Science (80- )* **235**: 1173–1178.
Williams KP, Ciafré S, Tocchini-Valentini GP. 1995. Selection of novel Mg(2+)-dependent self-cleaving ribozymes. *EMBO J* **14**: 4551–7.
Wilson DS, Szostak JW. 1999. IN VITRO SELECTION OF FUNCTIONAL NUCLEIC ACIDS. *Annu Rev Biochem* **68**: 611–647.
Wilusz JE. 2015. Removing roadblocks to deep sequencing of modified RNAs. *Nat Methods* **12**: 821–822.
Woodson SA. 2005. Metal ions and RNA folding: A highly charged topic with a dynamic future. *Curr Opin Chem Biol* **9**: 104–109.
Xu J, Appel B, Balke D, Wichert C, Müller S. 2014. RNA aminoacylation mediated by sequential action of two ribozymes and a nonactivated amino acid. *Chembiochem* **15**: 1200–9.
Yakhnin A V. 2013. A model for the origin of life through rearrangements among prebiotic phosphodiester polymers. *Orig Life Evol Biosph* **43**: 39–47.
Yang W, Lee JY, Nowotny M. 2006. Making and Breaking Nucleic Acids: Two-Mg2+-Ion Catalysis and Substrate Specificity. *Mol Cell* **22**: 5–13.
Yarus M. 2005. Chemical biology: Bring them back alive. *Nature* **438**: 2005.
Yarus M. 2011. Getting past the RNA world: the initial Darwinian ancestor. *Cold Spring Harb Perspect Biol* **3**: 1–8.
Yarus M, Widmann JJ, Knight R. 2009. RNA-amino acid binding: a stereochemical era for the genetic code. *J Mol Evol* **69**: 406–29.
Young KJ, Gill F, Grasby JA. 1997. Metal Ions Play a Passive Role in the Hairpin Ribozyme Catalysed Reaction. *Nucleic Acids Res* **25**: 3760.
Yu Z, Cao K, Tischler T, Stolle CA, Santani AB. 2014. Mung Bean Nuclease Treatment Increases Capture Specificity of Microdroplet-PCR Based Targeted DNA Enrichment. *PLoS One* **9**: e103491.
Zaborske JM, Narasimhan J, Jiang L, Wek SA, Dittmar KA, Freimoser F, Pan T, Wek RC. 2009. Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. *J Biol Chem* **284**: 25254–25267.

Zaher HS, Unrau PJ. 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13**: 1017–26.

Zhang B, Cech TR. 1997. Peptide bond formation by in vitro selected ribozymes. *Nature* **390**: 96–100.

Zhang B, Cech TR. 1998. Peptidyl-transferase ribozymes: trans reactions, structural characterization and ribosomal RNA-like features. *Chem Biol* **5**: 539–53.

Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput transfer RNA sequencing. *Nat Methods* **12**: 835–837.

Zhou W, Reines D, Doetsch PW. 1995. T7 RNA polymerase bypass of large gaps on the template strand reveals a critical role of the nontemplate strand in elongation. *Cell* **82**: 577–85.

Zhou XL, Du DH, Tan M, Lei HY, Ruan LL, Eriani G, Wang ED. 2011. Role of tRNA amino acid-accepting end in aminoacylation and its quality control. *Nucleic Acids Res* **39**: 8857–8868.

Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. 2012. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res* **40**: e54.

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

China Scholarship Council www.csc.edu.cn

**Titre :** Développement d'une méthode SELEX pour l'identification de ribozymes pour l'aminoacylation et analyse d'ARN aminoacylés dans le transcriptome d'*Escherichia coli*
**Mots clés :** ribozyme, SELEX, aminoacylation, tRNA, RNA-seq

**Résumé:** Les ribozymes sont des ARN naturels ou artificiels possédant une activité catalytique. Les ribozymes artificiels ont été identifiés in vitro par la méthode SELEX, et plusieurs d'entre eux ont été caractérisés par des études cinétiques. Ces molécules sont impliquées dans des réactions de clivage, de ligation, de modification d'extrémités d'ARN, de polymérisation, de phosphorylation et d'activation de groupements acyl. Parce qu'elle est nécessaire à la traduction, l'aminoacylation des ARN joue un rôle évolutif important dans la transition du monde de l'ARN vers le monde moderne de l'ADN et des protéines, et elle est centrale à l'établissement du code génétique. Plusieurs ribozymes catalysant le transfert d'acides aminés à partir de cofacteurs activants ont pu être isolés et caractérisés depuis une vingtaine d'années, ce qui a documenté la possibilité d'aminoacylation d'ARNt en l'absence des aminoacyl ARNt synthétases.

En développant un nouveau protocol SELEX basé sur l'oxydation au périodate, le but de notre travail est de découvrir de nouveau ribozymes d'une taille de l'ordre d'une vingtaine de nucléotides pouvant combiner la catalyse de l'activation des acides aminé et la transestérification. Bien que des molécules catalysant l'une ou l'autre des deux réactions ont été identifiées, aucun ribozyme n'existe à ce jour qui puisse utiliser des acides aminés libres et un cofacteur activant pour réaliser l'aminoacylation en 3' dans un même milieu réactionnel.

La sélection de molécules actives dans une approche SELEX exige la présence de régions constantes sur les deux extrémités des séquences pools aléatoires initiaux. Ces régions sont nécessaires pour l'amplification par PCR, mais elles imposent des contraintes importantes pour l'identification de ribozymes car elles peuvent complètement inhiber leur activité par interférence structurelle. Nous présentons un protocol optimisé qui minimise la taille de ces régions constantes. D'autre part, notre nouveau design est très spécifique pour la sélection d'ARN aminoacylés sur l'extrémité 3'. Ce protocol a été utilisé pour réaliser 6 à 7 cycles de sélection avec différents pools, et un enrichissement en séquences spécifiques a pu être mis en évidence. Bien que certains tests avec les pools sélectionnés a révélé une activité possible, des essais avec des séquences spécifiques de ces pools n'ont pour l'instant pas pu confirmer l'activité catalytique recherchée.

Un protocol basé sur le même principe de sélection a été utilisé dans une étude parallèle pour identifier les ARN aminoacylés présents dans l'ARN total d'Escherichia coli. Dans ce deuxième travail, note but est d'identifier tous les d'ARN aminoacylés par séquençage massif, avec à la clé la découverte possible de molécules autres que les ARNt et ARNtm. En utilisant les ARNt comme modèle, nous nous sommes aperçus qu'un protocol RNAseq standard n'était pas adapté à cause des bases modifiées présentes sur ces molécules. Nous avons développé et mis au point un nouveau protocol pour l'identification de n'importe quelle séquence aminoacylée en 3'. La nouvelle approche présentée devrait permette l'étude exhaustive de l'aminoacylation de toutes les séquences présentes dans l'ARN total.

**Title :** Development of a SELEX method to uncover auto-aminoacylating ribozymes and analysis of aminoacyl RNA from *Escherichia coli* transcriptomes
**Keywords :** ribozyme, SELEX, aminoacylation, tRNA, RNA-seq

**Abstract:** Ribozyme is a class of catalytic RNA molecule. Artificial ribozymes have been investigated by *in vitro* SELEX experiments and characterized by kinetic assays. Ribozyme is involved in RNA cleavage, ligation, capping, polymerization, and phosphorylation. Especially, RNA aminoacylation plays an important role in the evolution from the late RNA world to the modern DNA/protein world, and is central to the genetic code. Several ribozymes catalyzing amino acid transfer from various activating groups have already been selected in the past two decades, documenting the possibility of RNA aminoacylation with absence of aminoacyl tRNA synthetase.

With newly designed SELEX protocol, we aim to uncover small ribozymes of the order of 20 nucleotides that could catalyze both amino acid activation and transesterification. Although molecules catalyzing either reaction have been identified, no existing ribozyme could use free amino acids and activating cofactor(s) as substrates for 3' esterification in a single reactional context.

The selection of active molecules in a SELEX procedure requires the presence of constant tracks on both ends of the sequences constituing the initial random pools. These tracks are required for PCR, but they impose the burden to the identification of ribozymes. We present an optimized protocol that significantly minimizes the size of these constant tracks. And our newly design protocol is very specific for the selection of 3'-end aminoacylated RNA. We performed 6 to 7 cycles of selection with different pools, and observed an enrichement with specific sequences. Although some experiments did reveal a possible activity, no activity could be so far confirmed with specific sequences.

A similar protocol was also applied in a parallel study to identify aminoacyl RNA from total RNA in *E.coli*. Our goal is to possibly identify new classes of aminoacyl RNA by deep sequencing. Using tRNA to validate our protocol, we realized that a standard RNAseq procedure couldn't work due to the presence of modified bases. We established a new method for bank preparation to identify any sequence aminoacylated at the 3' end. Ultimately, this new approach will allow us to study the level of aminoacylation of any sequence present in total RNA.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France