Discovery and structure-function studies of key factors behind the non-canonical ZTGC-DNA observed in Siphoviridae family

Dariusz Czernecki

Sorbonne Université

Ecole doctorale Complexité du Vivant

Laboratoire Dynamique Structurale des Macromolécules à l'Institut Pasteur

Discovery and structure-function studies of key factors behind the non-canonical ZTGC-DNA observed in *Siphoviridae* family

Par Dariusz Czernecki

Thèse de doctorat de biophysique

Dirigée par Dr Marc Delarue

Présentée et soutenue publiquement le 14/12/2020

Devant un jury composé de :

Mme Mirjam Czjzek – Directrice de recherche CNRS – Examinatrice

M. Marc Delarue – Directeur de recherche CNRS – Directeur de thèse

Mme Marie-Agnès Petit – Directrice de recherche INRA – Examinatrice

M. Paulo Tavares – Directeur de recherche CNRS – Rapporteur

M. Eric Westhof – Professeur émérite – Président

Mme Sophie Zinn-Justin – Directrice de recherche CEA - Rapportrice

Abstract (English)

The subject for this thesis is to dissect the enzymatic pathway allowing a non-canonical base 2-aminoadenine, or diaminopurine (Z) to replace adenine (A) in the genomes of a number of *Siphoviridae* bacteriophages. 2-aminoadenine and thymine (T) form the Z:T pair bound by fully saturated triple hydrogen bond. Together with the standard G:C pair they form ZTGC-DNA which is resistant to host's restriction enzymes[1]. Although a biosynthesis pathway of diaminopurine in these phages was recently proposed (Sleiman *et al*., submitted), the reason for adenine's absence in phages' DNA was still a mystery.

I first focus on cyanophage S-2L, the originally-described bearer of 2-aminoadenine[2]. I identify a DNA primase-polymerase, PrimPol, responsible for its replication, with surprisingly similar activity towards dATP and dZTP. This prompted the characterization of a dATP-specific triphosphatase, DatZ. Its activity and conservation between the phages of interest explains the mechanism behind adenine removal. Secondly, I find that PurZ of phage S-2L's, a key enzyme in diaminopurine production, is not only an ATPase but also a dATPase. I identify a nucleotide pyrophosphatase, MazZ, as an essential component of the conserved Z biosynthetic pathway, that converts dGTP into dGMP, thus generating one of the substrates of PurZ. High resolution crystallographic structures of all 4 enzymes with their respective ligands explain the specificities observed in catalytic tests - or lack thereof.

Finally, I characterized the structure of a Z-specific family A DNA polymerase, PolZ, found in a related vibriophage ϕVC8 but absent in S-2L's genome. Its crystallographic structure in polymerase-exonuclease "coupled-open" and "coupled-close" states offers an explanation for the observed specificity. The complete description of 2-aminoadenine pathway could be subjected for bioengineering purposes.

Résumé (Français)

Le but de cette thèse est de décrire le chemin métabolique permettant de remplacer l'adénine (A) par la 2-aminoadénine (Z) dans le génome de bactériophages *Siphoviridae*. La 2-aminoadénine et la thymine (T) forment la paire Z:T, liée par trois liaisons hydrogène. Avec la paire G:C classique, ils forment un ADN « ZTGC » qui est résistant aux enzymes de réstriction de l'hôte[1]. Bien que le chemin de biosynthèse de la diaminopurine a été récemment proposé (Sleiman *et al*., soumis), l'absence d'adénine dans l'ADN de ces phages reste inexpliqué.

En premier lieu, mes travaux se sont focalisés sur le cyanophage S-2L, décrit comme porteur de 2-aminoadénine[2]. J'ai d'abord étudié la primase-polymérase, PrimPol, responsable de la réplication de l'ADN chez ces phages, et qui possède la capacité surprenante d'incorporer à la fois le dATP et le dZTP. J'ai ensuite caractérisé une triphosphatase dATP-spécifique, appelée DatZ, conservée chez tous les bactériophages *Siphoviridae*, responsable de la dégradation spécifique du dATP. J'ai également mis en évidence que PurZ, l'enzyme-clé pour la production de la diaminopurine, est non seulement une ATPase mais aussi une dATPase. J'ai identifié une nucléotide pyrophosphatase, appelée MazZ, composant essentiel du chemin de biosynthèse de Z, qui convertit dGTP en dGMP, en générant ainsi un des substrats de PurZ. Structures cristallographiques de haute résolution de tous les 4 enzymes avec leurs ligands respectifs expliquent les spécificités observées dans les tests catalytiques – ou leur absence.

Enfin, j'ai caractérisé une structure d'une ADN polymérase Z-spécifique de famille A, PolZ, trouvée dans le vibriophage ϕVC8, mais absente dans le génome de S-2L. J'ai résolu sa structure cristallographique en modes polymérase et exonuléase « couplé-ouvert » et « couplé-fermé », permettant de decrire son activité au niveau atomique. La description complète du chemin métabolique de 2-aminoadénine pourrait être soumise à la bio-ingénierie.

1. Szekeres, M. & Matveyev, A. V. Cleavage and sequence recognition of 2,6-diaminopurine-containing DNA by site-specific endonucleases. *FEBS Letters* **222**, 89–94 (1987).
2. Kirnos, M. D., Khudyakov, I. Y., Alexandrushkina, N. I. & Vanyushin, B. F. 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **270**, 369 (1977).

CONTENTS

INTRODUCTION 3

0. Exordium 4

I. Foundations of Life 5

1. The engine of life 5

2. Nucleotides and nucleic acids 7

3. Central dogma: the storage and flow of information 10

4. Translation to proteins 12

5. Metabolism of nucleic acids 15

A. Primases 16

B. Polymerases 17

C. Models of replication 19

II. Viral World 22

1. Viruses: emergence and classification 22

2. Structure and life cycle 24

3. Genomic organization 26

III. Departure from Universality 28

1. Natural exceptions in nucleic acid chemistry 28

2. Viruses with modified genomes 29

3. Artificial nucleotides and nucleic acids: going xeno 31

PROBLEM STATEMENT 34

1. Cyanophage S-2L and its ZTGC-DNA 35

2. Research goals 36

RESULTS 37

I. How cyanophage S-2L selects an alien base in its DNA: a structure-function approach 38

1. Preface 38

2. Article #1 43

II. Complete metabolic pathway of 2-aminoadenine biosynthesis in a clade of *Siphoviridae* phages 44

1. Preface 44

2. Article #2 48

III. Structure-function of the Z-selective DNA polymerase from bacteriophage ϕVC8 49

1. Introduction 49

2. Crystallographic methods 50

3. Description of ϕVC8 PolZ structure 51
4. ϕVC8 PolZ functional assays 56
PERSPECTIVES 61
I. General discussion 62
1. 2-aminoadenine metabolic pathways 62
2. DNA replication: ϕVC8 PolZ and PolA enzymes 62
3. Phage S-2L Z metabolism: the Z-cluster 64
4. Engineering of PurZ – alternative nucleic acid alphabets 65
II. Introduction of the S-2L Z-cluster into other species 67
1. Introduction 67
2. Z-cluster in *E. coli* 68
3. Z-cluster in phage T7 71
4. Conclusion 74
ANNEXES 75
I. PolA clustering and new subfamilies 76
1. Introduction 76
2. Results 78
3. Discussion 80
II. Fast and efficient purification of SARS-CoV-2 RNA dependent RNA polymerase complex expressed in *Escherichia coli* 81
1. Preface 81
2. Article #3 82
BIBLIOGRAPHY 83

INTRODUCTION

0. Exordium

The present thesis focuses on natural variants of DNA, their information-storing properties and their metabolism, and more specifically, on a set of viral exceptions in natural genomic DNA composition. To get the reader properly acquainted with the subject, a very broad introduction is provided, divided into three parts. Firstly, I discuss the universal biochemical basis of the living, driving the attention towards information storage and the basic metabolism of nucleic acids involved. Secondly, I introduce viruses and describe their basic properties, including interactions with their hosts and their replication mechanisms. Finally, in an attempt to challenge the canonical four-nucleotide (ATGC) hegemony, I describe the fascinating richness of known non-standard nucleic acids, both natural and man-made. I analyze their purpose and processing, concentrating especially on the viral examples: thus, I establish the niche in which my project is fitted.

Fundamental, well-established biological facts presented in the following work stay unreferenced. The reader may find them in molecular biology or molecular virology textbooks, such as:

- *Lehninger Principles of Biochemistry* (1),
- *Harper's Illustrated Biochemistry* (2),
- *Fundamentals of molecular virology* (3).

I. Foundations of Life

1. The engine of life

The word *biology* derives from Ancient Greek *βίος*- (bíos), meaning "life" or “lifetime”, and *–λογία* (-logía), meaning "study of". The term in its modern sense surfaced relatively recently, at the turn of the 19th century. Its first written definition appeared in 1799 in a work of Thomas Beddoes, an English physician:

Physiology therefore – or more strictly biology – by which I mean the doctrine of the living system in all its states… (4)

According to the word’s etymology, biological sciences are usually defined as dealing with the processes of life. The *Encyclopedia Britannica* gives the following definition of it:

Biology, study of living things and their vital processes. The field deals with all the physicochemical aspects of life. (5)

Another definition from the well-known journal of natural sciences, Nature, is the following:

Biological sciences encompass all the divisions of natural sciences examining various aspects of vital processes. The concept includes anatomy, physiology, cell biology, biochemistry and biophysics, and covers all organisms from microorganisms, animals to plants. (6)

As seen above, all definitions agree that biology refers to the science of the living. However, as usual, the devil lies in the details; here, an important issue is the meaning of the keyword *life*. Although usually referring to the self-sustaining, reproductive processes, the common (cellular) sense should perhaps be reconsidered in the scope of the recent advances in the field. Indeed, it is still frequently disputed, even if the dispute is purely academic, if the definition of life should be stretched to include the least complicated biological entities that live “on the border of life”; perhaps one could call them *the least of these my brethren* (7). The problem mainly consists of what is considered an environment, and how small a complexity of a biological entity would be allowed. In a nutshell, life is a social process and these social interactions make its very definition complicated.

First, aside from well-known “autonomous” cells (although no cell is truly autonomous from its environment), there exist obligate intracellular parasites, like eukaryotic protozoa (*Leishmania*,

Trypanosoma, *Plasmodium*, and *Toxoplasma*), fungi (*Cryptococcus neoformans*) as well as bacteria (*Chlamydia*, *Rickettsia* and *Coxiella burnetii*) (8). On their own, they are unable to conduct some parts of the necessary metabolic reactions, being forced to steal a number of pre-made metabolites from their host cells, such as ATP. Secondly, there exist viruses, which are even more primitive but work on a similar basis, taking advantage of a bigger portion of the host cell apparatus. The border between the two became even foggier with the recently discovered giant viruses, that carry in their genome a substantial portion of metabolism-related genes, as well as part of the translational machinery (aminoacyl-tRNA synthetases) (9), previously thought to be a hallmark of cellular life.

However, if one indeed stretches the definition of life up to viruses, should not the same be done with plasmids? Those short, naked DNA fragments live inside cells from all 3 domains of life – eukarya, archaea and bacteria – providing only genetic information, with no additional structural support (10). As a matter of fact, some plasmids are able to spread outside a cell in membrane vesicles (11) and homology studies point to occasional recombination between some plasmids and viruses (12). Much like plasmids, in eukaryotes there exist entire chromosomes, called *B chromosomes*, that are completely facultative: while they may affect host's fitness, their presence is seemingly not necessary for survival of anything but themselves (13). Let us go even deeper into simpler systems: the so-called *transposable elements* (or *transposons*) behave much like plasmids, but need to find a host DNA to insert themselves into and propagate with it. (14) Finally, there are *prions*: proteins that can force other, identical proteins with different structural conformation to assume their own fold (15). They can be regarded as a form of simple replication, dependent on a very specific environment.

Regarding those discoveries, one should perhaps rather speak about a *spectrum* of life, ranging from least complicated chemical machineries to man. In fact, as for their biochemical basis, even the simplest forms do not escape the universality of life's rules on Earth. The cellular life is constituted from 4 main classes of chemical components, aside from the predominant water solvent in which they bathe. More so, a large number of viruses incorporate all of them in their viral particle. These are: nucleic acids, proteins, lipids and glycans (16, 17). Their principal characteristics, that enable them to play distinct roles, are pointed out in Table 1.

	Chemical component			
	Nucleic acids	**Proteins**	**Lipids**	**Glycans**
Composition	C, N, O, H, P (rarely S)	C, N, O, S, H (rarely Se)	C, O, H, P	C, O, H, N
Monomer	nucleotide	amino acid	fatty acid	monosaccharide
Atoms in monomer (typically)	34-38	10-27	48-62	23-35
Monomers in molecule	2 000 to millions	several (peptide) to 35 000	2	several to dozens, thousands (polysaccharides)
Branching	no	no	not applicable	yes
Main character	acidic	various (all types)	hydrophobic	hydrophilic, acidic
Role	Information storage, regulation of expression, energy carrier (nucleotide), alarmone (nucleotide)	Catalysts, structural support (*e.g.* cytoskeleton, capsid)	Intra- and extracellular membranes, energy storage	Cellular recognition, cell wall, energy storage, protein folding

Table 1. Main characteristics of life's ubiquitous chemical components.

2. Nucleotides and nucleic acids

The structure of a nucleotide – the building block of nucleic acids – is tripartite, and perhaps chemically the most complicated of the four basic monomers of life (Fig. 1).

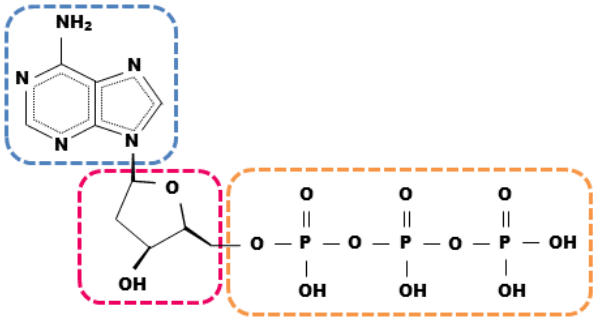


Figure 1. Chemical structure of a dATP nucleotide. Colored dashed rectangles delimit distinct moieties:

nitrogenous nucleobase (blue), deoxyribose sugar (magenta) and triphosphate tail (orange).

Firstly, there is a nucleobase, which defines the information content of its position in the polymer. It also promotes the stacking of the nucleic acid wire, contributing to its rigidity (18). In canonical cases, this base is represented by one in the set of four bases for a given nucleic acid. This set consists of two purines: adenine (A) and guanine (G), and two pyrimidines: cytosine (C) and thymine (T) for DNA or its close analogue, uracil (U), for RNA. The usual Watson-Crick base pairs (A:T/U and G:C) are formed through a mutual interaction of functional groups on the Watson-Crick edges; however, more sophisticated nucleic acids folds use other edges as well (Fig. 2).

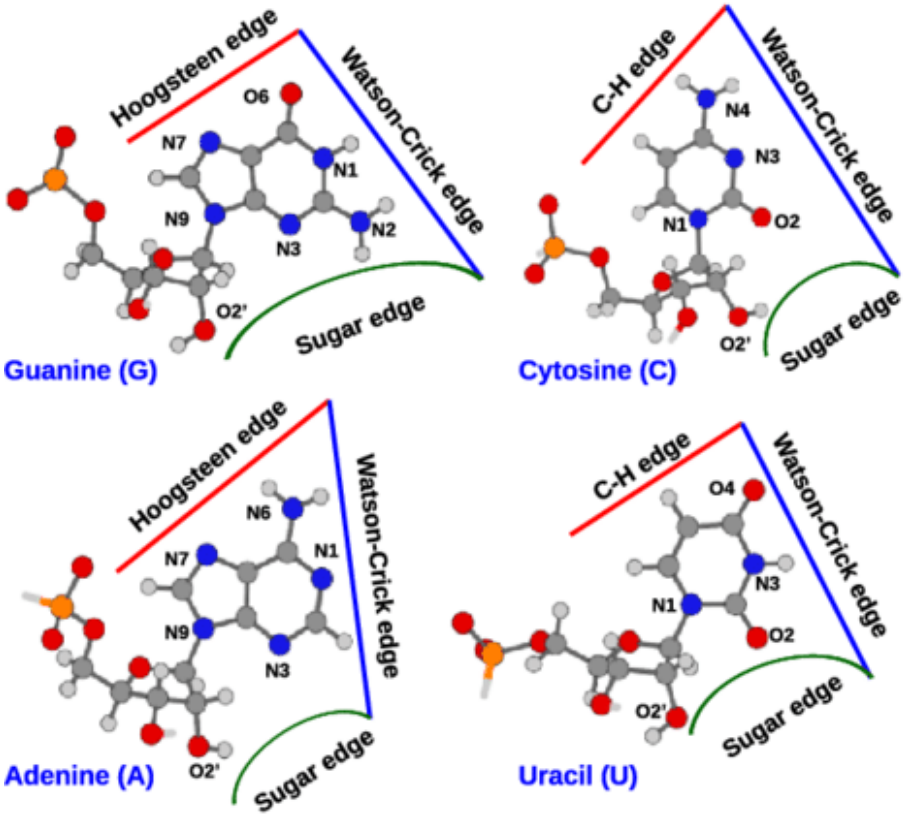


Figure 2. Edges of four standard ribonucleotides with an outline of their numbering on crucial heteroatoms, adapted from (19). Note position 7 of purines (Hoogsteen edge) and position 5 of pyrimidines (C-H edge).

Secondly, a nucleobase is joined with a pentose sugar, namely D-ribose; such compounds are called *nucleosides*. Because of the presence of this saccharide, nucleic acids actually overlap with glycans, and may be viewed as their highly specialized form. The presence or absence of a hydroxyl (-OH) group in position 2' determines if the chemical moiety in question is a *ribonucleoside* or *deoxyribonucleoside*, respectively. Nucleic acids composed of the former or the latter kind of monomers differ in flexibility (20), stability (21) and ability to fold into complex 3D-shapes (22, 23).

Thirdly, a nucleoside is attached to one or multiple phosphate groups, usually through esterification with the hydroxyl group on the 5' carbon atom, forming a *nucleotide* (ribo- or deoxyribo-). Covalent bonding between subsequent electronegative phosphate groups (numbered α, β, γ…) requires relatively high amounts of energy, which may be released when necessary.

Free nucleotides which are about to become incorporated into nucleic acids have three phosphates, and are called NTPs (*Nucleotide TriPhosphates*) or dNTPs, in case of deoxyribonucleotides. Curiously, it was shown that two phosphate groups have sufficient energy for the reaction of incorporation at higher temperatures (24), but this solution was not selected by nature. Additionally, adenosine triphosphate (ATP), or, more rarely, guanosine triphosphate (GTP) serve as a universal energy carrier, binding to many a protein catalyst that necessitate a supplementary driving force. Yet another nucleotide, a multi-phosphorylated guanine – ppGpp or pppGpp – serves as an alarmone in bacteria, changing promoter preference, which alters the transcription of certain genes and halts replication during nutritional stress (25).

Polymers of 5'-monophosphorylated nucleotides, bound together by binding their 5' α-phosphates to the 3' oxygen atoms of neighboring sugar residues are called nucleic acids (NA). Thus, their termini (if the molecules in question are not cyclic) have chemically different *5'*- and *3'-ends*, the former being phosphorylated. Everything towards the 3'-end relative to the point of interest is *downstream* of it, whereas the opposite (5'-end) direction is called *upstream*.

Depending on the kind of monomers they contain, nucleic acids are distinguished between RNA (*RiboNucleic Acid*) or DNA (*DeoxyriboNucleic Acid*), and can be *single*- or *double-stranded*. The iconic double-stranded form of DNA (dsDNA) comes from hybridization of two complementary strands, in an antiparallel fashion: the 5'-end is found opposite to the 3'-end of the second strand, and *vice versa*. This directionality holds true for dsRNA as well. Such double-stranded nucleic acids can assume typically 3 distinct helical conformations: A-, B- and Z-form, with nucleotides stacked inside the spiraling phosphosugar wires; nevertheless, other and rarer forms are possible as well (26–28). They differ by a wide range of structural parameters, including handedness, width and angles between nucleotides. Whichever form is assumed is usually determined by the kind of nucleic acid in question, its constitutive monomers and the solution the molecule is in; dsDNA usually assumes the B-form with distinct minor and major grooves (Fig. 3) while hetero-duplex RNA-DNA or dsRNA are usually in the A-form.

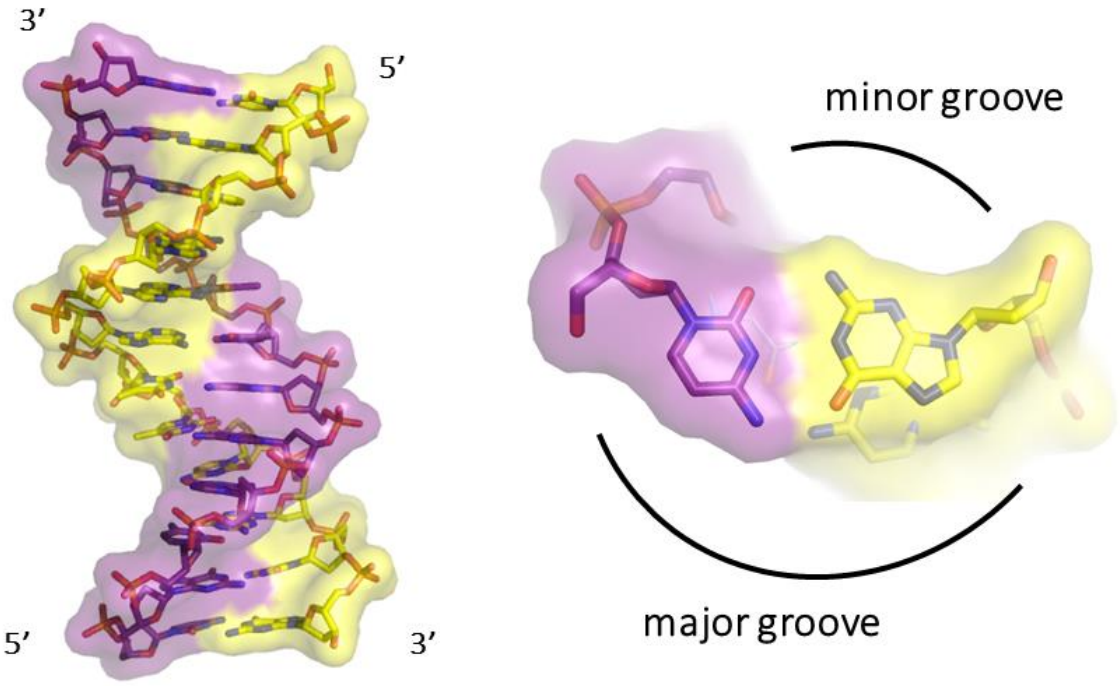


Figure 3. Right-handed B-DNA (PDB ID: 1BNA), viewed from the side (left) and bottom, down the helical axis (right). The two strands are colored in purple and yellow. The major groove can be seen facing the reader at the center (left) and below the base pair (right); it exposes to the solvent the 5-position of the pyrimidine and the 7-position of the purine. The less exposed minor groove is found on the opposite side of the base pair.

As a rule, the basic function of nucleic acids across all agents of life is information storage on carriers called *chromosomes* or facultative smaller plasmids, both circular or linear (mainly DNA), and regulation of expression during transcription and translation (RNA). An exception to this rule and perhaps the most innovative usage for DNA is exhibited by neutrophils, the immune cells, that eject their intracellular contents like a net, with DNA as a major component, to trap pathogens (29).

3. Central dogma: the storage and flow of information

The word *genetic*, describing hereditary information, is an adjective that comes from the noun *genesis*, and ultimately from Greek *γίγνομαι* (gígnomai): "I come into being" (30, 31). This word designates the source of information that the living matter execute: this genetic information is passed between generations of the same species (vertically) and directly between the existing organisms, not necessarily related (horizontally). It can be duplicated, extended and modified in a variety of ways, but is rarely spontaneously created *de novo*, as it is extremely unlikely that a carrier of valuable information can appear with the simultaneous ability to interpret itself (32). It is believed by many that this rule was violated only once, at the very dawn of life, creating an ancestor of all.

Only 63 years ago, in 1957, this cornerstone of biological knowledge was described in detail by Francis Crick. It sets the flow of information in biological systems between nucleic acids and from nucleic acids to proteins (33). It defined DNA and RNA as the ubiquitous carriers of information, and proteins, being their *expression*, as its effectors, performers: the latter's form, and thus the function, is determined by the sequence of the former through a process of *translation*. As the idea was confirmed to be universal to Earth, Crick called the idea a *central dogma*. As a matter of fact, although the definition of a dogma reads:

> *A principle or set of principles laid down by an authority as incontrovertibly true.* (34)

Crick later stated on multiple occasions that he ultimately regretted using the word, whose real meaning he did not know at the time (35, 36). The colleague of Crick, and his partner in the discovery of DNA's double helical structure, James Watson, later refined the central dogma, specifying that the information flows from DNA to RNA (*transcription*) and from RNA to proteins (*translation*) (37). That view, however, later collided with two exceptions.

The first one came with the discovery of reverse transcriptase (RT) in 1970 (38). This enzyme is capable of rewriting the information stored in single-stranded RNA (ssRNA) into complementary DNA (cDNA). It is found in retroviruses and retrotransposons (39) (here, the prefix *retro-* derives from the notion of *reverse transcription* (40)), but also in eukaryotic cells, where it plays the function of telomerase (41). The second exception is the surprising fact that the ribosomes, that carry out translation from RNA to proteins, are actually capable of translating a messenger DNA: although as of today it was proven only *in vitro* with cellular extracts – bacterial (42) and archaeal (43) – along with the additional use of antibiotics to enhance the DNA specificity, the usage of DNA as an immediate substrate for translation remains an open possibility.

Some argue that the existence of prions violates the original central dogma as well (44). While in all technicality it consists of a transfer of information between proteins, the nature of information is much more primitive and orthogonal (unlinked) to the information encoded in the genome. In my personal view, a protein kinase cascade, involving a chain of proteins that activate each other by phosphorylation, also passes some kind of information that is not directly encoded in the genetic material. Whatever the opinion, two facts always hold true: nucleic acids are the indisputable kings of information storage, and they cannot be directly reverse-translated from proteins – at least, not by any molecular machinery presently known, whether natural or artificial (Fig. 4).

Figure 4. Modern view of the central dogma of molecular biology. Typical flow of information in blue, atypical – in magenta. Debated protein-protein information transfer is shown by a dashed arrow. Note the absence of any arrow from protein back to RNA.

Finally, regarding the form of the nucleic acid carrying information – be it linear or circular, single- or double-stranded, one or multiple molecules, DNA or RNA – almost all combinations have been already observed in nature. The interpretation of the messages they encode constitutes the whole living domain: both material, like cells, and immaterial, like behavior (32).

4. Translation to proteins

Proteins, the most abundant components of a cell (45), are polypeptides made of simple but varied amino acids (aa). Their blueprint is contained in a *gene* made (usually) of DNA that is first transcribed by a protein called *RNA polymerase* into a *messenger RNA* (mRNA). This message is in turn interpreted by a protein-RNA macromolecular complex or machine called a *ribosome*. Nucleic acids strands with a translatable message from 5' to 3' are called *sense* (plus), and their complementary strands – *antisense* (minus). Thus, a specific sequence of (+)ssRNA is translated into a specific sequence of a protein, taking 3 nucleotides for 1 amino acid. This is mediated by the RNA-compatible carriers of amino acids, *transfer RNAs* (tRNAs), which decode one nucleotide triplet at a time.

Such encoding enables 64 different combinations of triplets taken from a 4-nucleotide set. In practice there are less than 64 unique tRNA classes in a cell: there are about 41-55 tRNA species expressed across eleven eukaryotes (46) and around 20 in mitochondria (47). In contrast, they are extremely abundant at the gene level, with every class having multiple copies: in humans, 497 nuclear tRNA genes were identified (48) (Fig. 5). Furthermore, they code for only 20-23 amino acids, along with the information to halt the translation of the mRNA through three *STOP codons*.

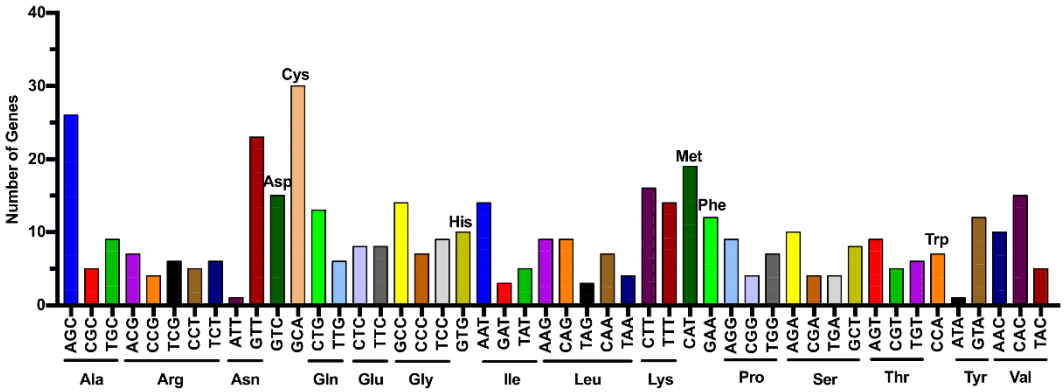


Figure 5. Distribution of human tRNA genes by the number of copies, according to their amino acid specificities (adapted from (49)).

Amino acids consist of a common backbone, having – according to their name – both a carboxyl and an amino functional groups attached to a central carbon atom, called *carbon alpha* (C_α). Each C_α is also linked to one hydrogen atom and the side chain of an amino acid (R) specific to its type; their subsequent carbon atoms follow the nomenclature (C_β, C_γ…). Four different groups on C_α enable the molecule to have two distinct *enantiomers* (mirror conformations). Proteogenic amino acids follow an L- conformation under the archaic D-L notation (Fig. 6): as a consequence, all functional groups in a protein protrude to the same direction relative to the chain, or *backbone*.

Figure 6. L- (left) and D- (right) conformations of an amino acid. Amino group in orange, carboxyl group in purple; group *R* in green designates a side chain.

A protein is created by the condensation of the carboxyl group of one monomer and the amino group of the following monomer, creating a planar *peptide bond* (Fig. 7). Thus, the backbone consists of the C_α atoms and the peptide bonds. Like nucleic acids, acyclic proteins have chemically different termini: after the symbols of atoms involved in these groups, the extremities of a polypeptide chain are called the *N-terminus* (the beginning) and *C-terminus* (the end).

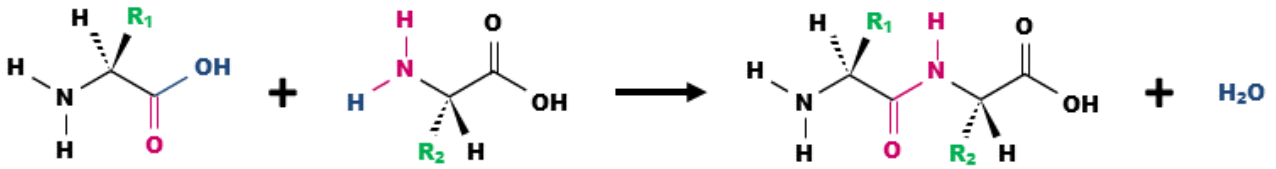


Figure 7. Condensation reaction between two amino acids. Atoms in magenta contribute to the peptide bond, and lie in one plane; atoms in blue form a water molecule. Amino acid rests are in green.

There are four structural levels to a protein. The primary structure concerns the sequence of its *residues*, from N- to C-terminus. Typical proteins use the standard set of 20 amino acids, but some of them may be modified post-translationally (50). In addition, there exist three more proteinogenic amino acids encoded already by tRNA molecules of some organisms. Those are: selenocysteine, present in every domain of life but not all lineages, sometimes inserted after a special mRNA structural signal; pyrrolysine, occurring in some prokaryotes, that replaces one STOP codon (51); and formylmethionine, used to initiate protein synthesis (START codon, identical to the Met codon) in bacteria (52) and organelles (53), but which may be removed post-translationally (54).

The secondary protein structure of a protein describes which parts fold into helices (α, 3_{10} and π, differing by the number of residues per turn), β-strands making planar β-sheets, and short β-turns (55). The ternary structure is the three-dimensional fold that an independent *domain* of a protein adopts, arising from the interactions between secondary-structure elements and individual residues, guided by long-range (sequence-wise) contacts. It usually results in hydrophobic residues residing in the core of a protein, while exposing polar amino acids to the solvent. Finally, proteins often interact with one another, and the quaternary structure describes the *ensemble* of all proteins involved in a functional complex. Whereas the protein's sequence and its interaction with the immediate environment usually lead to its folding in a determined and unique manner, there also exist intrinsically disordered proteins or regions thereof, that are also biologically active (56).

With the exception of information storage, proteins play all kinds of roles in a biological entity. They can be of structural importance to a cell or a virus, act as a regulator or inhibitor of gene expression, or be a reaction catalyst (an *enzyme*).

5. Metabolism of nucleic acids

There is a number of proteins that interact with nucleic acids, nucleotides and their precursors, to ensure their proper processing. In bacteria, under optimal condition and during stress, roughly 2-3% of proteins participate in nucleotide metabolism, and further 2-4% in DNA transactions such as replication, recombination, topology control and repair (57, 58); these numbers increase by ~1% for the synthetic minimal bacterial genome of JCV-syn3.0 (59).

The nucleobases and their nucleotide derivatives are generated by numerous nucleotide *synthetases*, that is *synthases* (enzymes catalyzing a synthesis reaction) that use other nucleotides (ATP or GTP) as a source of energy, and *lyases*, that degrade their substrates. As it happens, because of high similarity between the words "synthase" and "synthetase" resulting in frequent mistakes, the Joint Commission on Biochemical Nomenclature proposed the replacement of the latter by *ligases* (60); however, in retrospect, it only added to the confusion, as old names were partially retained. The corresponding triphosphates are formed and recycled in a salvage pathway by *kinases* and *phosphatases*, catalyzing opposite reactions of phosphorylation and dephosphorylation, respectively.

The key feature ensuring the propagation of genetic material is the ability to duplicate it. This process is divided into two stages: *priming*, which begins the replication on a single strand, generating short *primers*; and *polymerization*, during which the following nucleotides are added in a complementary fashion using the short double-stranded region as an anchor point. Thus, proteins involved in these activities are called *primases* and *polymerases*. They are intimately related to each other; the clear distinction between them sometimes disappears, when in specific situations they are taking over each other's function.

The recurring molecular mechanism behind the activity of many a protein involved in nucleic acid metabolism is a *two-metal-ion mechanism*. It involves, accordingly to the name, two divalent metal ions that, on one hand, neutralize the negative charge of an involved phosphate, and, on the other hand, stabilize a nucleophile (a hydroxide ion, OH^-) opposite to the bond of the free or leaving group that is about to be hydrolyzed (61) (Fig. 8). It is found in primases (62), polymerases (63), nucleases (61), phosphatases (64) and even catalytic RNA (*ribozymes*) processing themselves or other RNA molecules (65).

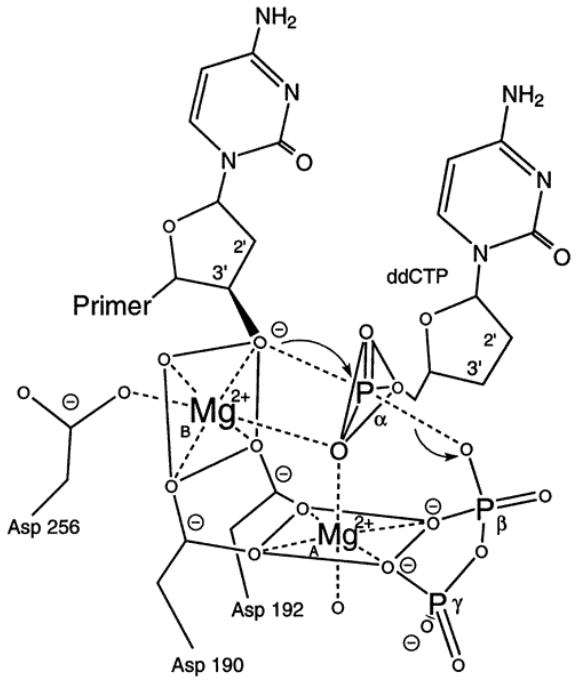


Figure 8. Nucleotidyl transfer mechanism in human Pol β using the two-metal-ion mechanism (adapted from (66)). The activated 3'-OH group of the last nucleotide of the primer strand (left) attacks the alpha phosphate of the incoming nucleotide (right), causing the departure of the pyrophosphate. Note the octahedral geometry of the coordination spheres of both Mg^{2+} ions, sharing one phosphate oxygen as an axial ligand.

A. Primases

There exist two superfamilies of primases, originally distinguished between bacterial and archaeo-eukaryotic ones, although recent progress has unraveled the presence of the former in phages, and of the latter in all instances of life, including mobile genetic elements, viruses and plasmids (67). They are called DnaG and AEP (Archaeo-Eukaryotic Primase), respectively, but both usually consist of three domains playing particular roles (Fig. 9).

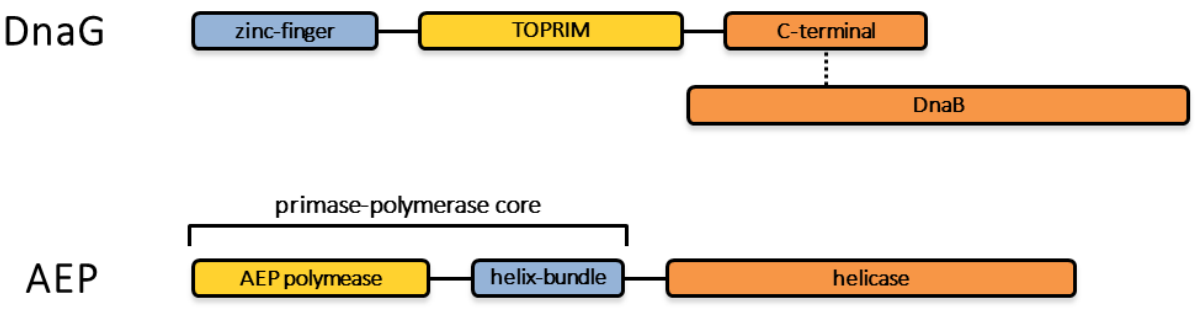


Figure 9. Schematic domain structure representation of typical DnaG and AEP enzymes. Their three domains

hold primase activity (yellow), confer sequence specificity (blue) or act as a helicase (orange).

Firstly, a region of DNA upon which they would act has to be unwound from its double-stranded form by a helicase. In bacteria, this is the responsibility of a partner protein, DnaB, which interacts with DnaG through the C-terminal domain of the latter (68); in contrast, a helicase is often directly fused as a C-terminal domain in AEP (69). Secondly, a specific sequence in such a non-hybridized region is targeted. Such is the role of the N-terminal zinc-binding domain of DnaG (70), as well as helix-bundle domains (PriCT-1, PriCT-2, PriL or PriX) that constitute the middle domain of a typical AEP protein (67). Thirdly, a special polymerase domain, with the proper primase activity, adds the first few nucleotides in front of the instructing strand, using (d)NTPs as a substrate. It is represented by the middle domain of DnaG, the TOPRIM (topoisomerase-primase) domain (71), or the N-terminal AEP domain itself (the name of the superfamily is derived from the domain's name) (69). Interestingly, the AEP domain may also carry a zinc-finger (sub)domain within itself, although it is not present in all lineages (72).

Most frequently, priming involves the synthesis of short RNA fragments which are subsequently removed by DNA polymerases with 5'-3' exonuclease activity (73) or specific nucleases such as ribonuclease (RNase) H (74). Nevertheless, some AEP primases called PrimPol have double primase-polymerase activity, and can also use dNTPs directly for priming (75, 76). Extraordinarily, in a specific replication mode of some plasmids and viruses discussed further below, a DNA-dependent RNA polymerase (RNAP) was shown to produce the primers, instead of either DnaG or AEP (77).

Finally, the need of a nucleic acid primer may be bypassed altogether. Some B family DNA-dependent DNA polymerases (pPolB) can interact, through a fused TPR1 domain, with a terminal protein (TP) covalently bound to linear plasmids, transposons and viral chromosomes, called *invertrons.* They use a hydroxyl group of an exposed TP's serine residue to begin the DNA synthesis (78, 79). Other, related polymerases from the same family (piPolB) do not even need the terminal protein for priming (80).

B. Polymerases

Compared to only two main types of primases, polymerases (Pol) are more diverse. They can operate on different kind of single-stranded template, being *DNA*- or *RNA-dependent*; similarly, the nature of their product dictates if the enzyme in question is a *DNA* or *RNA polymerase*. No distinction is made after the primers they use, as in general they accept both kinds. Additionally, some polymerases are able to extend the nucleic acid chain without taking instruction from any template at all: they are

called nucleotidyltransferases or terminal transferases (81, 82).

There seems to be only one superfamily for RNA-dependent RNA polymerases (RdRP) (83). Their counterparts, DNA-dependent RNA polymerases mentioned in the previous section (often abbreviated to RNA polymerases; RNAP), are globally represented by evolutionarily distinct single- and multi-subunit forms (ssRNAP and msRNAP) (84, 85), although there exist instances of single-subunit variants related exclusively to msRNAP (86). In contrast, DNA polymerases divide into three distinct folds, and can be further distributed between 7 families (87): 6 of them are DNA-dependent (A, B, C, D, X, Y), while 1 represents the RNA-dependent reverse transcriptase (RT). The summary of distribution between folds and families of all DNA polymerases, along with related AEP superfamily, are shown in table 2.

	Polymerases		
Folds	Klenow	Polβ	Two-barrels
Families	A, B, Y, RT, AEP, RdRP, viral ssRNAP, plasmidic ssRNAP, mitochondrial ssRNAP	C, X	D, all msRNAP, cellular ssRNAP

Table 2. Repartition of known polymerases according to folds and families (86–89).

Different polymerases were selected as main replicators for each domain or form of life (87). These are: family C for bacteria; family B for eukaryotes; family B or D for archaea (90); family A for organelles (91, 92); family A, C or AEP for plasmids (93–95); family A, B, C or AEP for DNA viruses (76, 96); family RdRP for RNA viruses (97); and family RT in concert with RNAP for retroviruses and retrotransposons. The remaining two families (X and Y) are specialized in DNA repair functions.

Polymerases always add a nucleotide by attaching the 5’ α-phosphate of an incoming nucleotide (n+1) to the 3' hydroxyl group of the previous one (n); thus, polymerization occurs in the 5'-3' direction. As stated before, the two-metal-ion mechanism (Fig. 8) is at the heart of such a transferase activity, although recent research shows evidence for a transient third divalent metal ion, at least in families X and Y (98, 99).

Polymerases from families A, B, C and D are fused to (PolA, PolB, some PolC) or bind (some PolC, PolD) a proofreading exonuclease domain (87). It cleaves the newly synthetized chain in an opposite, 3’-5’

direction, removing incorrectly inserted nucleotides and lowering the intrinsic error rate of a DNA polymerase, 1 per 10^4-10^5 nt, by 2-3 orders of magnitude (100). Additionally, some family A polymerases, like *E. coli* Pol I, fuse with a 5'-3' exonuclease able to excise RNA primers in front of the polymerase (upstream on the template strand).

C. Models of replication

Primases and polymerases are mandatory to replicate a genetic material. The whole process of nucleic acid duplication is however more complex, employing a plethora of proteins playing various functions. As there exist multiple replicative DNA polymerases, its details can vary greatly between the agents of life.

The region where replication is initiated is called the *origin of replication* (ori); there can be one or multiple origins per chromosome. In a typical dsDNA, the process starts with the assembly of the pre-replication complex at such sites. The complex recruits a helicase, that performs DNA *dehybridization*, or nucleic acid unwinding, advancing through dsDNA along the replication. Although it generates an overwinding of the helical structure downstream, this structural tension is progressively relieved by a topoisomerase. During the process, separate strands of DNA are kept apart by single-stranded DNA binding proteins (SSB).

Usually, replication progresses in both directions from the origin, resulting in two different *replisomes* on the so-called *replication forks* (Fig. 10). At these forks, a new DNA strand is synthetized on each ssDNA template: one following the direction of unwinding (occurring on the *leading* strand), and the other, assembled from smaller segments (*Okazaki fragments*, 150-200 nt) synthesized in the opposite direction (*lagging* strand). To begin, primases synthesize 10-30 nucleotide-long RNA primers: only one for the leading strand, and one for each Okazaki fragment. From there, DNA polymerases continue to manufacture complementary chains in the 5'-3' direction. They are often bound to other replication actors (such as helicases and primases) and processivity or remodeling factors (like DNA sliding clamp/PCNA or thioredoxin (101), respectively) in order to remain bound to DNA, or to stay *processive*. RNA primers are subsequently removed, either degraded by a 5'-3' exonuclease or displaced by the polymerase and cleaved off by a flap endonuclease (102), making room for the remaining DNA synthesis. Finally, the missing bounds between the ends of freshly synthetized DNA segments are joined by a DNA ligase. In case of linear chromosomes, their endings – repetitive sequences called *telomeres* – are reduced in length after each duplication due to the need of an

overhang template for a polymerase. They are either reconstructed by a telomerase complex, or simply reduced after each replication.

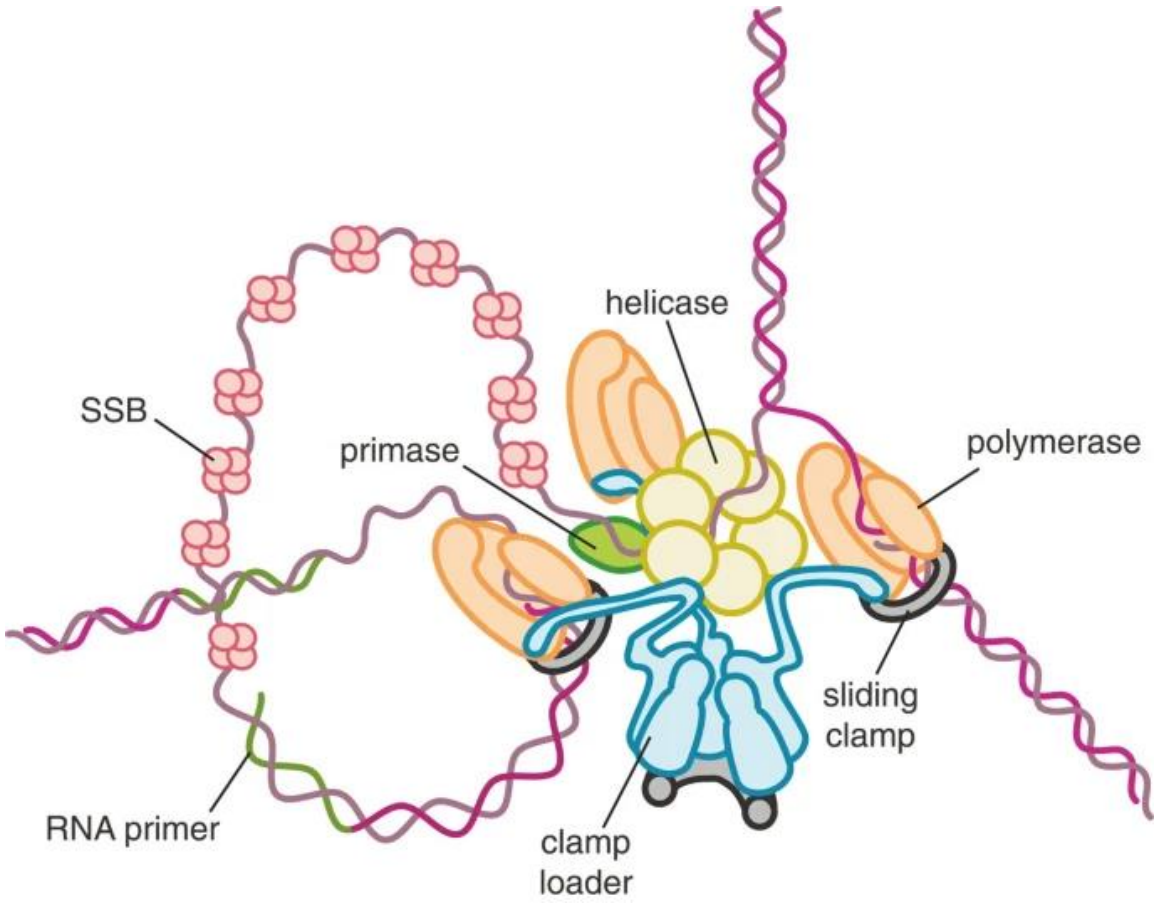


Figure 10. A schematic representation of a bacterial DNA replication fork, with main agents involved (adapted from (103)). A clamp loader places the sliding clamps onto dsDNA strands.

The above model of DNA duplication is known as symmetric replication, and is common to the three cellular domains of life. For circular molecules, it passes through a *theta structure*, named after the microscopic structure similar to Greek letter Θ (theta) visible during the process. In contrast, some of the simpler entities have adopted other solutions.

As an example, the duplication of mitochondrial DNA follows a strand-displacement model (104). In this asymmetric case, replication starts at the origin of heavy-strand synthesis (O_H). At first, only the leading strand is being duplicated, while unwinding the other one and keeping it single-stranded. Eventually, the origin of light-chain synthesis (O_L) is reached, from where the replication will start on the other strand in the opposite direction. Due to this delay, the second strand may also be replicated in a continuous fashion as a leading strand, as a large enough fragment of the dsDNA has already been dehybridized and made available for complementary synthesis.

Furthermore, some plasmids and viral chromosomes adopt the rolling circle (RC) model of replication. At the double-strand origin (DSO), dsDNA refolds into a cruciform structure similar to a Holliday junction (105). Then, a discontinuity (*nick*) is made on one strand by a Rep protein. A polymerase uses the free 3'-end to start the synthesis of, once again, only the leading strand, displacing the other during the process. After coming a full circle and replicating the whole leading strand, the displaced remainder of the second strand is cleaved off, and its loose ends fused anew, forming ssDNA. Finally, RNAP binds to the single-strand origin (SSO) of the second strand and creates an RNA primer (77); a DNA polymerase finishes the replication.

II. Viral World

1. Viruses: emergence and classification

As mentioned above, viruses turn out to be a contentious issue with respect to the definition of being *alive*, missing a huge chunk of metabolic machinery and at least one basic ingredient of molecular biology, namely the ribosome. Despite this severe deprivation, they come with assets of their own: some viruses can, for example, carry out reverse transcription and protein-based priming, unlike any other cellular automata. For that reason, the *virosphere* (concerning all viruses on Earth) is an interesting place to search for novel biochemical inventions.

Viruses are simple entities separate from, but dependent on the cellular tree of life. They consist of rogue nucleic acids – DNA or RNA – embedded in a proteinaceous bag. Viruses can invade cellular environments to use the wealth of their biological compounds and enzymes, that the former do not bother to (and cannot) synthesize. Although cells from all three domains can become a viral prey, for historical reasons a virus attacking bacteria is explicitly called a *bacteriophage* (or simply a *phage*) (106).

The origin of viruses is unknown; in fact, there could have been multiple distinct origins. Despite the lack of a precise scenario, the general consensus is that they are extremely ancient, and three hypotheses are usually considered to explain their emergence (107). The *virus-first* hypothesis states that they have emerged from the primordial soup before cellular life, guided by their simplicity and frequent lack of homology. On the other side, *degeneracy* and *vagrancy* hypotheses propose that they have arisen from a cell, that has either deteriorated, or has given up a small but autonomous part of its genome. Finally, some viruses also show high similarity with plasmids (11, 12) or even mobile elements (108), and a recent review considers their evolutionary relationship (109).

The first attempt to classify viruses that received broad attention was proposed in 1962 by A. Lwoff, E. Horne and P. Tournier (the LHT system): they distributed them mainly with respect to the kind of their nucleic acid, capsid symmetry and presence or absence of a membrane envelope (110). Following it, another impactful organization of the virosphere, called the *Baltimore classification*, was developed in 1971 by David Baltimore. He divided them into 7 groups, according to the exact nature of their genetic material and its intermediate forms before producing viable transcripts (Table 3).

Group	I	II	III	IV	V	VI	VII
Genome	dsDNA	(+)ssDNA	dsRNA	(+)ssRNA	(-)ssRNA	(+)ssRNA	dsDNA
Intermediates before mRNA	-	dsDNA	-	(-)ssRNA	-	(-)ssDNA, dsDNA	(+)ssRNA, (-)ssDNA, dsDNA

Table 3. Baltimore classes of viruses, with the corresponding genome composition and processing. Plus and minus signs denote sense and antisense strands, respectively.

Since the early 1970s, the International Committee on Taxonomy of Viruses (ICTV) is the body in charge of virus classification. In 1991 it devised a 5-rank structure, roughly matching the reduced hierarchical structure used for cellular organisms. In 2019 these rules were revisited, increasing to as much as 15 levels of taxa (111) (Fig. 11). The highest rank in the new taxonomy of viruses is called a *realm*, and corresponds to the cellular life's domain. On the other side of the classification, the lowest taxon *species* was retained. As the most recent classification appeared just before the time of writing the present thesis, many new ranks are as of yet unassigned or only postulated: for example, one proposed viral realm is *Riboviria*, including all RNA viruses bearing an RNA-dependent RNA polymerase.

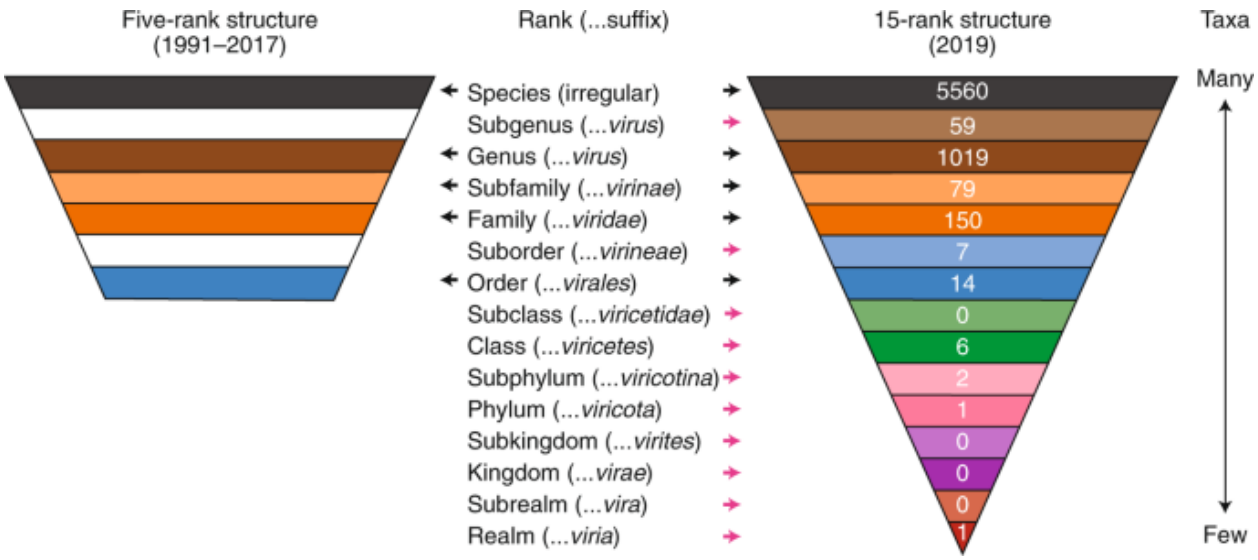


Figure 11. New 15-rank virus classification (adapted from (111)).

2. Structure and life cycle

Viral genomes travel from cell to cell in protective protein shells (*capsids*), most often arranged periodically in a symmetric structure, or closely approaching symmetry. Sometimes, they are additionally embedded in a lipid bilayer enriched with host's and viral proteins, constituting their *envelope*, which however turns to be fragile in most cases (112). The complete viral particles (*virions*) come in various shapes and sizes, but most frequently they have a 20-400 nm diameter, being about 10 times smaller than a typical bacterial cell. As for exceptional cases, viruses devoid of a protein capsid are called *viroids*, and those with a capsid stolen from other viruses during co-infection – *virusoids* (113). Even more bizarrely, scientists described an ssDNA virus encoding a major capsid protein found only in ssRNA viruses, proving possible recombination between distant classes (114).

Despite having a standalone form, viruses depend entirely on their cellular hosts to perform metabolic processes and reproduce. As such, the first stage of viral life cycle is their proper recognition. It is performed through the attachment of the receptors on viral envelope or parts of their capsids (especially tails) to the exposed host's proteins.

From there, viruses employ either one of 3 strategies for nucleic acid entry. Firstly, they can simply be ingested by the host through endocytosis, which dissolves their encasing and releases its contents; for instance, such is the case for SARS-CoV-2 (115). Secondly, the viral envelope may directly fuse with cellular membrane, like does HIV (116). Lastly, the virus can pierce cell's membrane and inject its nucleic acid into host's cytoplasm. This leaves the capsid outside of the cell, and is conducted by phage T7 (117) and other bacteriophages.

Once freed, viral genome hijacks cellular machinery, creating a *virocell* governed by the virus' instructions (118). It is done through transcription of the first portion of viral genes in the cytoplasm: some of the transcripts inhibit certain cellular processes, while others, especially viral enzymes, establish novel pathways. After the control is taken and the biochemical ground is set, viral genome can replicate. For DNA viruses infecting eukaryotes, this stage usually takes place in the nucleus, and RNA viruses prefer to reproduce in cytoplasm. However, both of these behaviors were found to have exceptions: for example, Influenza virus, a (-)ssRNA virus, replicates as ribonucleocapsids in the nucleus (119), while giant dsDNA viruses create replicase complexes (RC) in the cytoplasm (120).

While viral replicative proteins get in charge of replicating the nucleic acid, other proteins form the capsids. They either assemble around the new genomes directly, or come together to form hollow

shells, which are filled with nucleic acids during *packaging*. In the final stage, new virions burst out of the cell, sometimes departing with parts of its membrane containing both host and viral proteins which form new envelopes. Due to much smaller size than that of their host, replication of viruses is highly efficient: for phages T7 and λ, one virocell gives progeny of more than 100 viral particles in less than 25 min or more than 50 min, respectively (121, 122), but the numbers can go up to dozens of thousands for other viruses like SIV, a simian relative of HIV (123).

In an interesting turn of events, there exist viruses that depend not on regular cells, but on virocells themselves. While most of such *satellite viruses* rely on coinfecting helper viruses for replication, as they do not encode their own polymerase (124), some can inhibit their helpers, becoming thus *virophages*, like Sputnik preying on giant mimiviral virocells (125).

The above replication pathway is referred to as the *lytic cycle*, common to all types of viruses. In contrast, some species are able to integrate their genetic material into their host's DNA (Fig. 12). In this form of *provirus*, they remain inactive for multiple generations and replicate with the genome of the host cell. It is believed that viruses undergoing this pathway, called the *lysogenic cycle*, can occasionally mutate and lose the ability to subsequently reactivate. It results in "fossil" surplus DNA not directly involved in cellular processes; in humans, such proviral fossils make up to 8% of the genome (126).

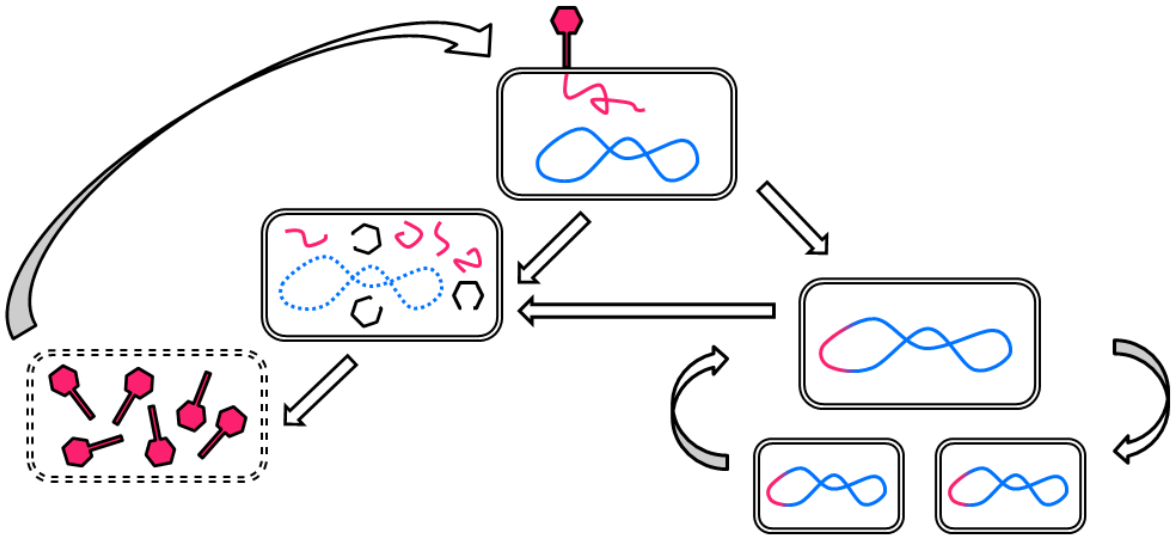


Figure 12. Intertwined lytic (left) and lysogenic (right) cycles of a virus. Cellular DNA is in blue, viral DNA in magenta.

3. Genomic organization

As it is the case for other chromosomes and plasmids, viral genomes may also be circular or linear. Usually, viral genomes are under evolutionary pressure for compactness (127). Their genes follow themselves very closely, frequently overlapping; often, multiple ones are condensed on singular, *polycistronic* mRNAs, or are even expressed together in *polyproteins*, that are subsequently cut up by endopeptidases. They can, however, bear introns, enabling alternative splicing. Viruses are attuned to their hosts in terms of nucleic acid composition (GC content) and codon usage, albeit rather loosely, while staying seemingly closer to their respective families in that respect (128–130).

The physical distribution of genes on the viral genome is seldom random. It tends to reflect their time of expression and function, and it can usually be divided into three distinct groups of class I (*early*), II (*middle*) and III (*late*) genes or proteins, like in the case of phage T7 (131) (Fig. 13).

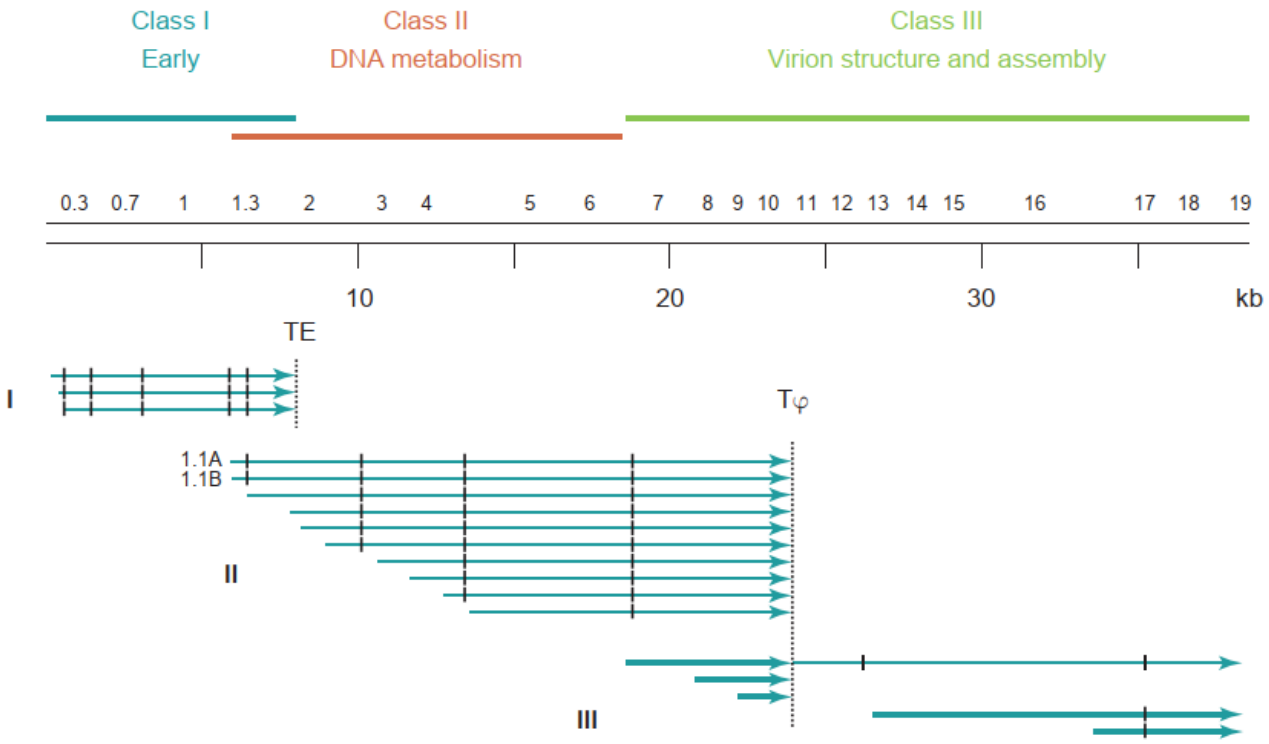


Figure 13. Phage T7 genetic (above) and transcriptional (below) maps (adapted from (3)). TE and Tϕ stand for transcriptional terminators.

Early genes are readily expressed by the host's machinery, producing inhibitors of cellular functions such as transcription and defense. In addition, they encode viral RNA polymerase or a specific σ factor that enables recognition of the subsequent viral promoters by the cellular RNAP. Interestingly,

the transcription of early genes can drive internalization of the rest of the genome (132). Middle proteins are involved in genome replication – as such, many enzymes processing nucleotides or nucleic acids are found in this region. Aside the ones participating in viral genome synthesis, one of them may be a nuclease degrading cellular nucleic acid, in order to supplement the nucleotide pool avidly consumed by the virus. Finally, late transcripts consist mainly of structural proteins forming the viral capsid, along with release enzymes such as lysozyme that degrades the cell wall.

III. Departure from Universality

1. Natural exceptions in nucleic acid chemistry

Both DNA and RNA are universal and found in all cellular domains of life. Additionally, viruses and smaller independent entities use one of these polymers to store their genetic information. The conservation involves not only the sugar moiety, deoxyribose or ribose, but also the set of four nucleobases, with the exception of DNA's thymine having a supplementary methyl group on the C5 atom of the pyrimidine ring compared to RNA's uracil.

Despite this unifying rule, many diverse biological entities found a way to extend their nucleotide repertoire with modified analogues. Modifications can occur on nucleotides before being incorporated into a nucleic acid (pre-replicative) or directly on polymers (post-replicative). Such alterations may be simple (unique), or multiple at different positions (hypermodifications). Finally, they can appear with random frequency or only in well-defined sequences: in the extreme case, in some simple viral entities, the modified nucleotides can completely replace the original ones.

Perhaps the most commonly known modifications are observed in rRNA and tRNA molecules participating in translation, involving among others methylation, acetylation, pseudouridylation or thiolation (133–135) (Fig. 14). This heavily conserved attunement influences folding and catalytic properties of ribozymes and reinforces the decoding accuracy of tRNA (136). Eukaryotic mRNA can be modified too: aside from 5' capping enhancing its stability (137), methylation of adenine in position 6 regulates protein expression (138), whereas acetylation of cytidine was found to promote the efficiency of translation (139).

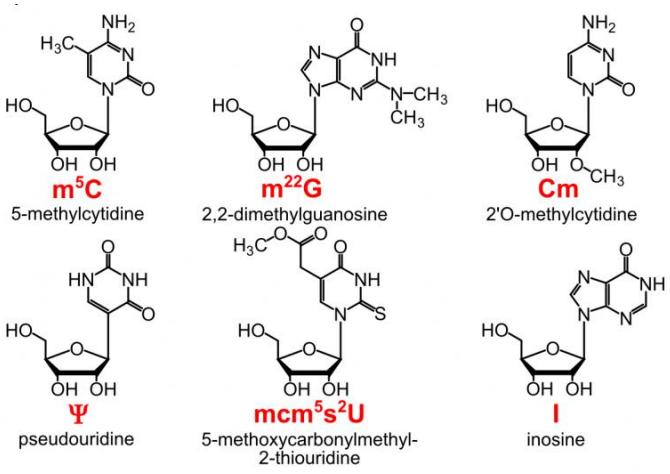


Figure 14. Typical RNA modifications in the base moiety (adapted from (140)).

On the other hand, modifications of the genomic material usually consist of additional groups protruding outside the nucleic acid double helix, in the major groove. In this way they are not interfering with base pairing, enabling compact information storage on buried, protected nucleobases. Instead, they influence interactions with DNA-binding proteins, allowing distinction between different kinds or parts of DNA through decoration of a common underlying core. If such modification is reversible, it results in a case of gene expression regulation known as *epigenetics*, from Greek *ἐπι-* (epi-), "over, outside of, around".

A frequent example of epigenetic control is mammalian 5-methylcytosine (5-mC) in CpG dinucleotides that can group in high-frequency *CpG islands*: their methylation usually silence underlying transcription promoters (141). Some methylcytosine residues can be further oxidized to 5-hydroxymethylcytosine (5hmC) found in mouse and human DNA (142). Finally, in *Trypanosoma brucei*, the hydroxyl group of hydroxymethylcytosine is a subject of β-D-glucosylation, forming base J at 5-20% frequency (143). In bacteria, methylation of genomic and plasmidic cytosine and adenine regulates gene expression and protects their DNA from being degraded by restriction-modification (RM) systems (144). Indeed, nucleobase modifications can be a direct consequence of RM modification modules, or, as shown below on phages, they are devised to avoid restriction modules targeting foreign DNA.

Lastly, modification of DNA is not limited to nucleobases: products of a bacterial gene cluster are able to incorporate sulfur into nucleic acid's backbone, generating phosphorothioate group (145). The modification is sequence-specific, but variable between species, with frequency less than 1.5% (146).

2. Viruses with modified genomes

Viral genetic modifications are believed to serve the purpose of escaping the host's defense systems, such as restriction-modification enzymes or the CRISPR-Cas system. To that end, coliphages T2, T4 and T6 employ 5-hydroxymethylcytosine, similarly to eukaryotes (147). Contrary to the latter, however, here the modification of cytosine is pre-replicative (at the level of triphosphate) and, because of simultaneous dCTP pool depletion (148), substitution of C for 5hmC in their DNA is complete. Most of 5hmC is additionally post-replicatively glucosylated, with varying efficiency between the T-even phages (149) (Fig. 15). In a similar vein, other bacteriophages attacking *Bacillus subtilis* use hydroxymethyluracil (5hmU) instead of thymine (150).

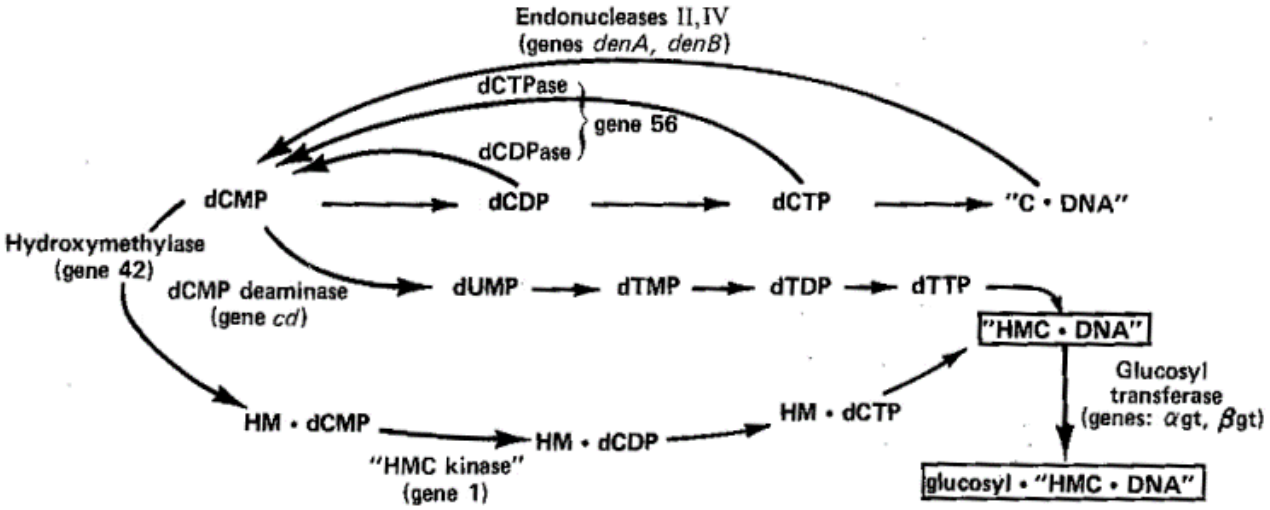


Figure 15. Pathway of (glucosyl-)hmC-DNA induced by T-even phage infection (adapted from (147)).

Another group of viruses, including coliophage 9g and possibly phage BRET, replace genomic guanine with archaeosine (G+): in a chain of pre-replicative enzymatic reactions involving QueC, QueD and QueE, guanine taken from GTP is transformed into 7-deazaguanine with a formamidino group attached to the C7 atom (151, 152). Interestingly, a ribonucleotide version of G+ was previously found in archaeal tRNAs, which gave the nucleobase its name (153). Recently, three more 7-deazaguanine modifications, intermediates in the G+ metabolic pathway – queuosine (ADG), $PreQ_0$ and $PreQ_1$ – were found in other phages, and previously in bacterial tRNA as well (154).

Lastly, an adenine-deprived cyanophage S-2L was found to replace the base with a close analogue, 2-aminoadenine (155): the special character of this substitution (Fig. 16) and its underlying metabolism is thoroughly investigated in later sections of this thesis.

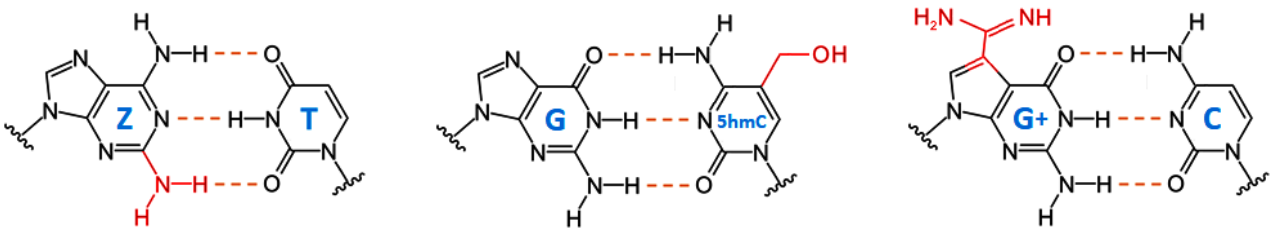


Figure 16. Viral non-standard genomic nucleobases: 2-aminoadenine paired with thymine (left), guanine paired with 5-hydroxymethylcytosine (middle), archaeosine paired with cytosine (right). Highlighted in red are modifications with respect to natural nucleobases.

Although not confirmed experimentally, *in silico* predictions propose novel enzymes involved in nucleobase modification and identify new biochemical pathways, many of them in viruses (156).

3. Artificial nucleotides and nucleic acids: going xeno

Inspired by natural solutions, human creativity has produced a number of unique nucleotides and their polymers. They are termed *xenonucleotides* and *xeno nucleic acids* (XNA), after Greek *xeno*, meaning "foreign" (157). In these molecules, all chemical moieties may be subject to modification: due to their tripartite nature, they can be divided into three groups, with hypermodified hybrids at their intersections (158) (Fig. 17).

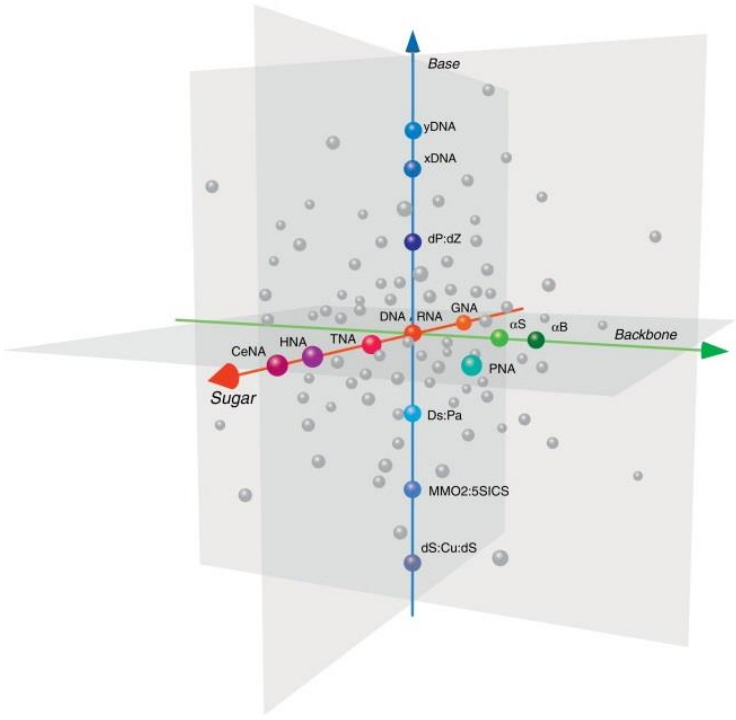


Figure 17. Distribution of XNA on three orthogonal arbitrary axes, corresponding to base, sugar and phosphodiester backbone modifications (adapted from (158)).

A special subset of sugar-modified nucleic acid are mirrored molecules, containing L-(deoxy)ribose: L-DNA and L-RNA (159, 160). They may constitute mirror DNAzymes or ribozymes, termed *Spiegelzymes* (from German *Spiegel-*, "mirror") (161) or aptamers (*Spiegelmers*) (162). Moreover, these polymers were successfully transcribed and reverse-transcribed by a mutant mirror enzyme, Y-family DNA polymerase Dpo4 (163). Polymerization trials are not limited to known, but mirrored sequences: multiple sugar-modified nucleic acids were amplified or reverse-transcribed to DNA with mutants of family B TgoT polymerase from *T. gorgonarius* (164) (Fig. 18).

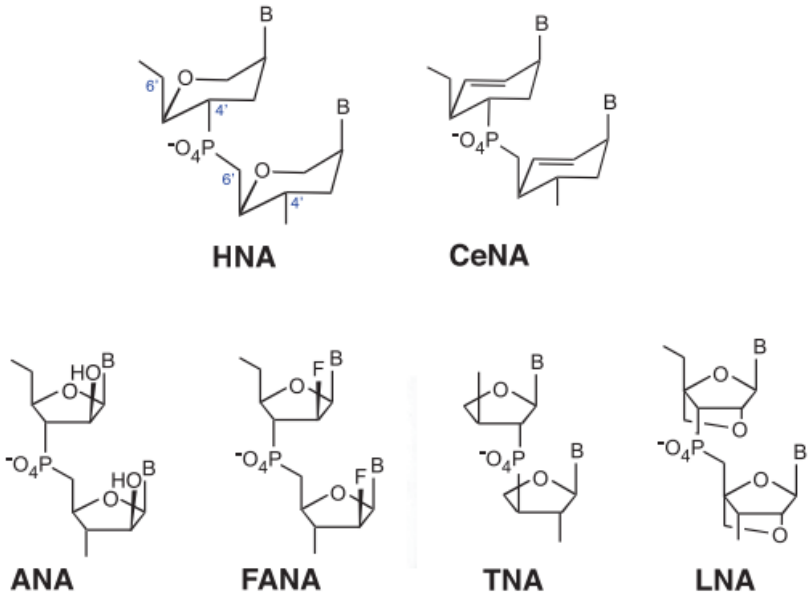


Figure 18. XNAs processed by mutants of family B DNA polymerase TgoT, selected by directed evolution (adapted from (164)). B stands for natural nucleobases.

Not only the subject of fundamental research, artificial nucleotides and nucleic acids have already found multiple applications. Acyclovir, an analogue of guanine nucleoside with incomplete sugar moiety, has pharmaceutical use against *Herpesvirus* family (165). It is converted in human cells by viral thymidine kinase and host kinases to a corresponding triphosphate, with inhibits viral DNA polymerases. On the other hand, Locked Nucleic Acid (LNA), an RNA analogue with an extra bond on the ribose limiting its conformational freedom, is used as a *small interfering RNA* (siRNA) silencer of complementary mRNA (166, 167).

One of the most prominent examples of successful xenonucleotide implementation into a living organism involves a single hydrophobic 5SICS-NaM (X:Y) base pair introduced into a *E. coli* plasmid (168) that stays stable after multiple generations, if the cell is externally provided with the corresponding triphosphates (169). Furthermore, the structural basis of incorporation of this and similar base pairs was investigated *in silico* and *in crystallo* for *T. aquaticus* DNA Pol I (KlenTaq) (170).

Finally, in the most recent approach, standard DNA and RNA alphabet was expanded by two synthetic base pairs, identical for both polymers save for a small variation for one nucleobase. These new base pairs that closely resemble natural ones result in *hachimoji* nucleic acids, from Japanese *hachi-* (“eight”) and *-moji* (“sign) (171) (Fig. 19). The structure of such DNA is not deformed with respect to the natural one, and is readily transcribed by a mutant of T7 RNA polymerase.

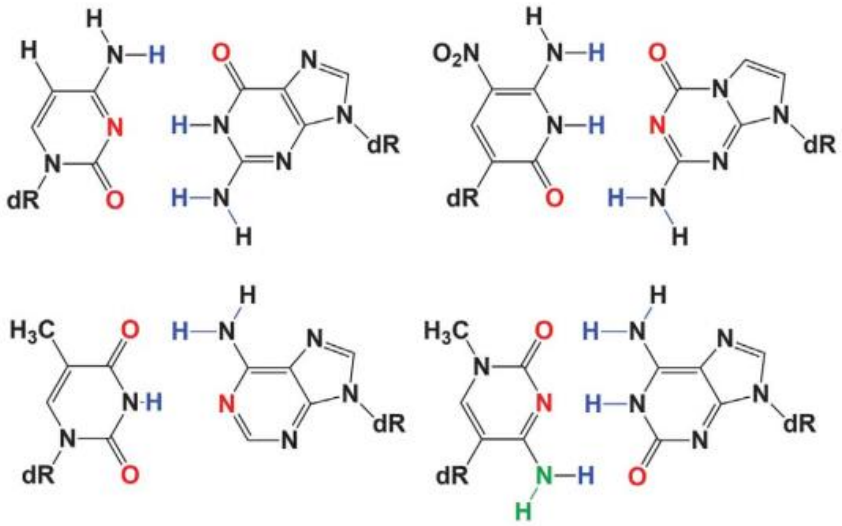


Figure 19. Hachimoji DNA nucleobases: two standard (left) and two synthetic (right) base pairs (adapted from (171)). dR stands for the remaining deoxynucleotide moieties. Hydrogen bond donor atoms are highlighted in blue, acceptor atoms – in red. The base that varies between DNA and RNA has an atypical amino group in position 2 that hinders transcription (green). Note the position of nitrogen atoms in the aromatic rings: they are not pyrimidines nor purines, but close analogues, except for isoguanine (bottom right).

PROBLEM STATEMENT

1. Cyanophage S-2L and its ZTGC-DNA

Synechococcus virus S-2L, a bacteriophage of cyanobacteria (cyanophage), was discovered in 1977 "*in the water samples from the outskirts of Leningrad* [present Saint Petersburg]" (155). It belongs to the *Siphoviridae* family (Realm: *Duplodnaviria*; Kingdom: *Heunggongvirae*; Phylum: *Uroviricota*; Class: *Caudoviricetes*; Order: *Caudovirales*), and has a typical morphology of a siphovirus, with a long tail crowned by the DNA-containing head (Fig. 20, left).

Characterization of S-2L's genetic material, dsDNA, revealed an unusual nucleobase composition: it included guanine, cytosine and thymine, but not adenine. Instead, its close analogue, with a supplementary amino group in position 2, pairs selectively with thymine (Fig. 19, right). This novel purine is known as 2-aminoadenine or 2,6-diaminopurine, denoted onwards by the letter Z. As the modification is found on the Watson-Crick edge, 2-aminoadenine is the first and, so far, the only natural genomic nucleobase analogue known to directly alter the interaction between two complementary nucleic acid strands.

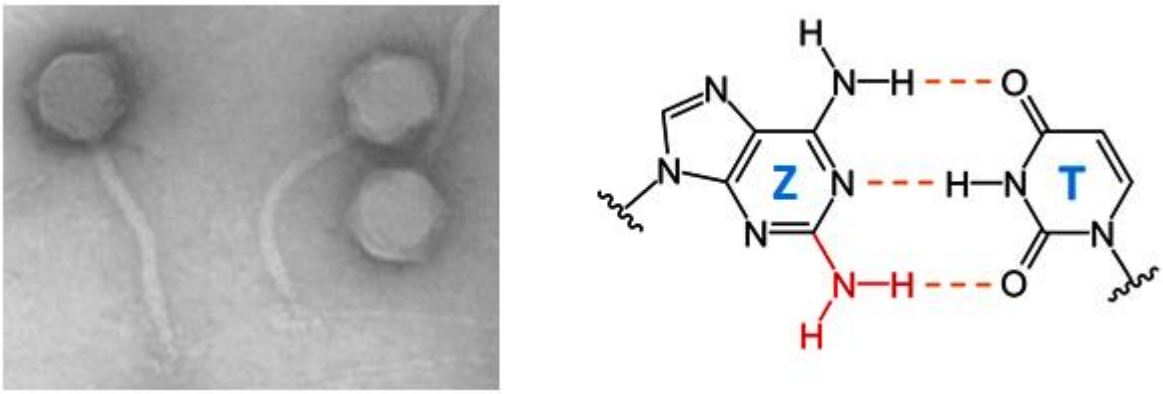


Figure 20. Cyanophage S-2L (left, adapted from (172)) and its 2-aminoadenine:thymine (Z:T) basepair (right). The supplementary amino group with respect to adenine is colored in red; hydrogen bonds are marked by orange dashed lines.

Because of the additional third hydrogen bond, the resulting ZTGC-DNA of cyanophage S-2L has a higher melting point than standard ATGC-DNA (172, 173). 2-aminoadenine presence alters the typical ultraviolet and CD absorption spectra of DNA; the base itself and its nucleotide are also fluorescent in physiological conditions, similarly to another close analogue, 2-aminopurine (172, 174, 175). Finally, the A-to-Z replacement makes the S-2L DNA unrecognizable by the host's defense mechanisms, namely adenine-targeting restriction enzymes (176).

Since its discovery the genome of S-2L has been sequenced, revealing an enzyme eventually confirmed to catalyze a key reaction in 2-aminoadenine nucleotide biosynthesis (Sleiman *et al*., submitted). However, despite the proposed pathway explaining its provenance, the question of how Z is selectively incorporated in front of T in the phage's genome – with a specific DNA polymerase or otherwise – remained unanswered.

2. Research goals

My interest in non-standard, but natural DNA prompted me to concentrate my research on diaminopurine metabolism and its DNA incorporation, under the direction of Marc Delarue in the Unit of Structural Dynamics of Macromolecules, in Pasteur Institute. The aim of my doctorate was tripartite:

- to fully characterize, both functionally and structurally, all enzymes of cyanophage S-2L involved in 2-aminoadenine selective incorporation into DNA during replication;

- to search for related phages with the identified key enzymes, compare their genomic organization and complete structure-function studies of S-2L diaminopurine metabolism;

- to participate in the description of Z-selective DNA polymerase found in newly discovered S-2L relatives, but not in S-2L, through structural studies on the enzyme from the related *Vibrio cholerae* phage (vibriophage) ϕVC8 (177). This family A polymerase (PolZ) has 5'-3' polymerase and 3'-5' exonuclease activities.

The structural part, one of the pillars of this work, makes use of X-ray crystallography, the technique with which I became familiar over the course of the project. Each of these three objectives culminated in a scientific article, either submitted or in the final stage of writing. I redacted the first two as a first author and the structural chapter of the third as an associated author. These three chapters-articles constitute the main body of the following part – the results of my research.

As a bonus, I re-assessed family A DNA polymerase diversity using a clustering method, discovering a new major PolA subfamily (Pol I D) and tracing the evolutionary history of PolZ. I have also participated in a work that describes efficient purification of an active three-subunit RNA-dependent RNA polymerase of the infamous SARS-CoV-2. Both items are disclosed in the *Annex* section, the latter in the form of an article submitted to PLoS One with me as an associated author. The former could serve as the basis of a subsequent article.

RESULTS

I. How cyanophage S-2L selects an alien base in its DNA: a structure-function approach

Czernecki D, Legrand P, Rosario S, Tekpinar M, Kaminski PA and Delarue M

submitted to *Nature Communications* (July 2020) – under revision

1. Preface

In the first article we unravel the mechanism by which 2-aminoadenine is selectively incorporated during cyanophage S-2L DNA replication, which had remained a mystery for more than 40 years.

Specifically, we identify a member of the PrimPol family as the sole possible polymerase in S-2L genome and demonstrate that it has both polymerase and primase activity, with a puzzling preference for A vs Z in front of a T. Its crystal structure at 1.5 Å resolution confirms there is no structural element in the active site that could lead to the rejection of A in front of T. Contrary to our expectations, the discovery of the PrimPol enzyme does not explain why the phage's DNA contains only Z and no A in front of T. To resolve this contradiction, we investigate the function of nearby genes and show that one of them is in fact a triphosphohydolase specific of dATP (DatZ) that leaves intact all other dNTPs, including dZTP. This profoundly modifies the pool of available dNTPs and explains the absence of A in S-2L genome: the strategy is reminiscent of what is observed in T-even phages. Crystal structures of DatZ at sub-angstrom resolution (0.86 Å) with various ligands allow to describe its mechanism as a typical two-metal-ion mechanism and to set the stage for its engineering.

Aside from the obvious biological importance of this work, explained in detail in the article's content, one of the protein of interest, DatZ, is also of particular crystallographic interest. During the screening process DatZ crystallized frequently, in multiple forms (Fig. 21), with crystals sometimes almost as big as the crystallization drop.

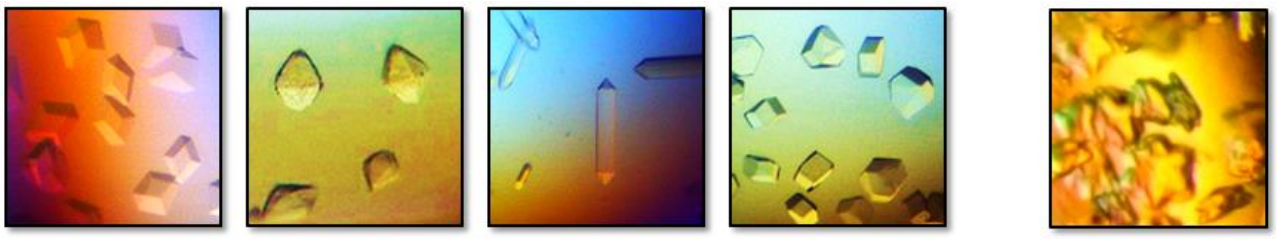


Figure 21. Multiple crystalline forms of DatZ (left). The rhombic ones (first image) were grown from a simple crystallizing solution (Table 4), were highly reproducible and almost always devoid of apparent crystal defects. They were chosen for solving the structure of all DatZ complexes presented here. On the right, an alternative, star-shaped crystal form of DatZ is shown, bound to dATP in EDTA.

The 0.86 Å DatZ structure resolved at synchrotron Soleil is presently the world's 14th best resolution structure for comparable proteins (between 150 and 200 aa). This was possible thanks to the hexametric quaternary structure of the enzyme that enabled crystallization in a compact R3 space group. At that level of resolution, the experimental $2F_o$-F_c electron density gave the detailed position not only for the vast majority of light atoms, but even for several hydrogen ones, especially well-defined in sulfhydryl group of S118 (Fig. 22).

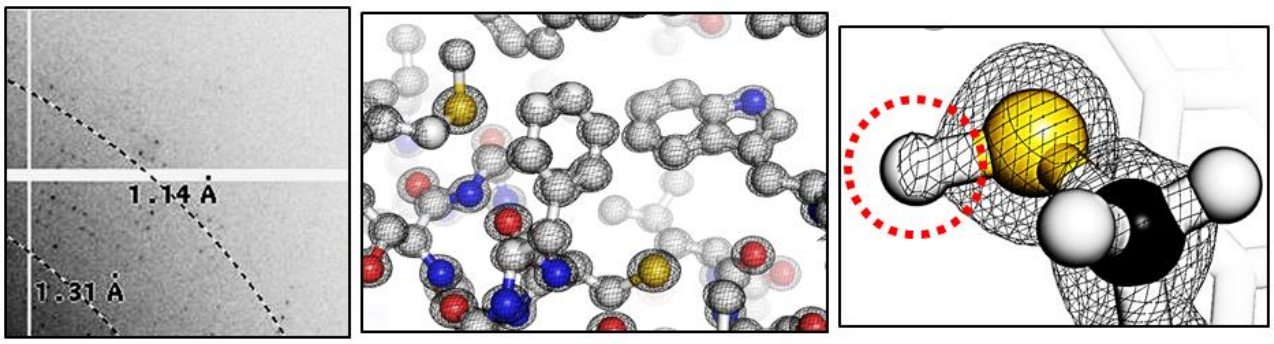


Figure 22. Diffraction pattern of DatZ collected at the synchrotron Soleil, with several diffraction spots around 1 Å visible to the naked eye (left). The electron density of the crystal DatZ structure (0.86 Å) contoured at 3 sigmas (black mesh) pin-points the position of individual atoms and is visibly proportional to their electron number (middle). At 1 sigma, the density of S118 sulfhydryl hydrogen is clearly visible (right; indicated by a red doted circle).

I also crystallized and obtained high-resolution crystallographic data for DatZ bound to dA, Co^{2+} ions and metavanadate (VO_3^-) or wolframate (tungstate, WO_4^{2-}) anions, mimicking the penta-coordinated transition state of the leaving phosphate. The presence of both anions between the two catalytic Co^{2+} cations was confirmed with anomalous signal (Fig. 23): however, they are rotated 77-81° away from

the position of the dATP's α-phosphate with respect to the Co^{2+} axis (3.3 Å away from P_α), extending to the open pocket in the direction of the γ-phosphate.

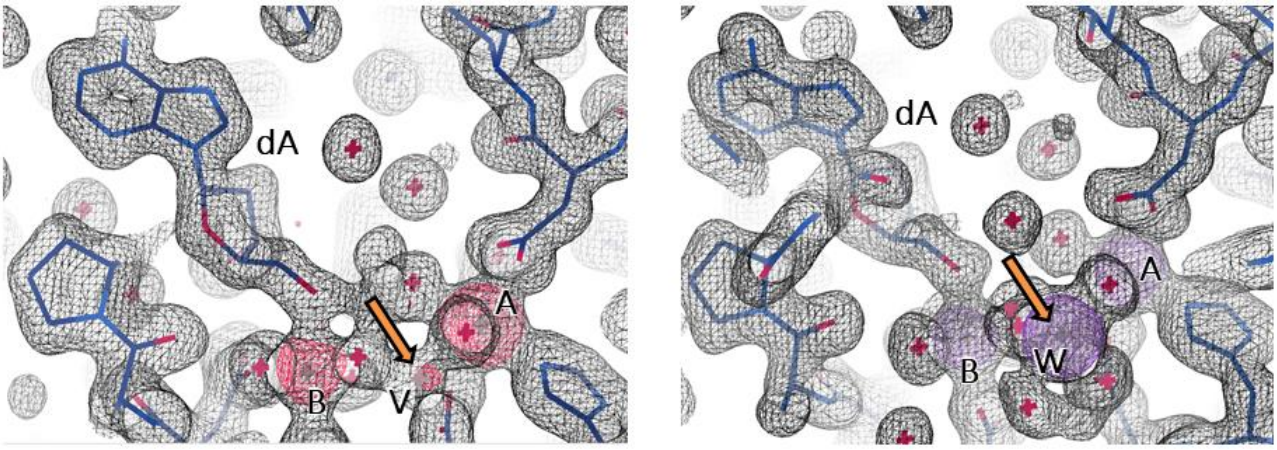


Figure 23. DatZ with bound dA, Co2+ catalytic cations A and B and metavanadate (1.44 Å, left) or wolframate anion (1.48 Å, right), unrefined data. Electron density is contoured at 1 sigma (black mesh). The anomalous signal collected at wavelength 0.9801 Å is contoured at 5 sigmas for both metavanadate (red) and wolframate (purple); the heavy atoms are indicated by orange arrows. Crystals were obtained by adding to the solution saturated $NaVO_3$ or 100 mM Na_2WO_4 in proportion ≃1:10.

The observed outcome could mean that the enzyme has high affinity for dATP, but it exerts a simultaneous force of ejection on the triphosphate, resulting in efficient evacuation of the product. In such a scenario, phosphate analogues are not able to penetrate deep enough to attain the position of P_α: instead, they stop at the next location in phosphate ejection process, with an apparently higher affinity for the anion. This behavior may be connected to the lack of activity observed with dADP and dAMP. Our hypothesis is that the enzyme binds these substrates correctly (similar K_d), but the phosphate tail is too short for proper orientation by residues K81 and K116, which results in the incorrect positioning of P_α for the reaction to take place (much lower k_{cat}).

For PrimPol, the full construct never crystallized during screens, presumably due to the allegedly flexible three-domain structure of the enzyme. Its two-domain derivative, PP-N300, gave only rare microcrystals. At first the mono-domain PP-N190 gave spiked precipitates ("urchins"), that were gradually optimized, up to visually perfect crystals. Unfortunately, these had an intrinsic default, being most probably twinned, preventing the resolution of the structure despite vast efforts from our part. The structure of PP-N190 was ultimately solved from a different crystal form, thin needles that grew together in easily separable beams, that arose in new crystallization conditions (Fig. 24).

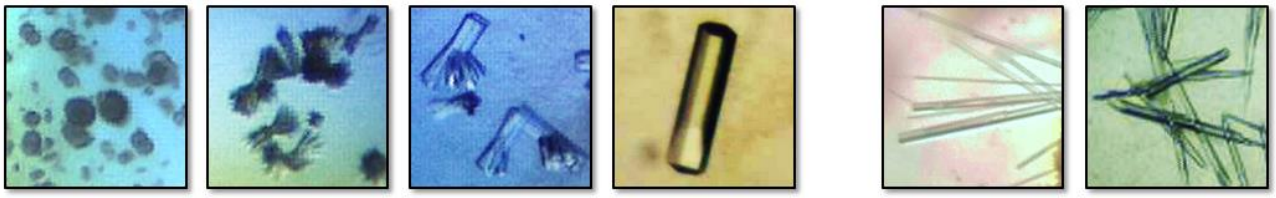


Figure 24. Gradual improvement of first, well-diffracting but defective PP-N190 crystals (left). The structure was eventually solved from another crystalline form (right).

The conditions for all the crystals mentioned above are gathered in Table 4.

Crystals	Conditions
PP-N190 (defective)	• $(NH_4)_2SO_4$ (2 M), PEG 400 (2% v/v), HEPES pH 7.5 (0.1 M), $NiCl_2$ (14 mM)
PP-N190 (used)	• $CaCl_2$ (0.1 M), isopropanol (5% v/v), PEG 8000 40% (20% w/v), MES pH 6 (0.1 M)
DatZ (rods)	• NaH_2PO_4 (1 M), K_2HPO_4 (0.66 M), Li_2SO_4 (0.2 M), CAPS pH 10-10.5 (0.1 M)
DatZ (used)	• Li_2SO_4 • Na formate (3-4 M) • $(NH_4)_2SO_4$ (1.5-2 M), optionally NaCl (0.2 M) all buffered in pH 7.0-8.5 (HEPES, TRIS 0.1 M)

Table 4. Crystallization conditions for PP-N190 and DatZ crystals. Each bullet point stands for one independent condition. All assays were done at 18°C, except the second one, done at 4°C.

Finally, while looking for diaminopurine specificity factors, I have expressed two more proteins of S-2L involved in DNA processing: VRR-NUC nuclease and exonuclease VIII. I tested their activity as endo- and exonuclease in a simple assay with closed and open (cut once with a restriction enzyme) standard RSF1-Duet plasmid (Fig. 25).

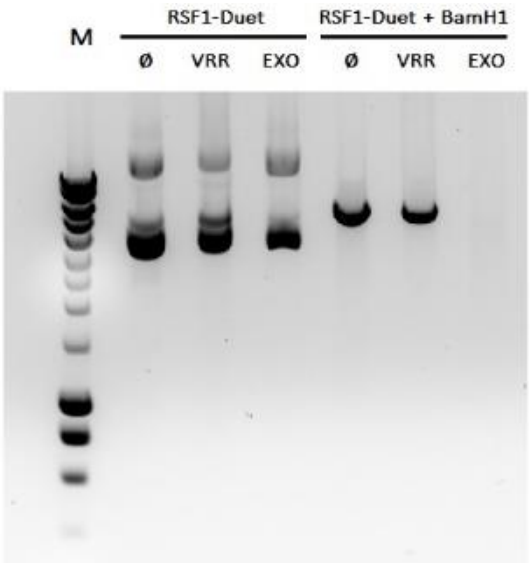


Figure 25. Endo- and exonuclease assay for S-2L VRR-NUC (VRR) and Exonuclease VIII (EXO) enzymes, visualized on 1% agarose gel. The test was made with 1.2 µg of closed and 0.3 µg of open RSF1-Duet plasmid in 5 mM $MgCl_2$ and 50 mM TRIS pH 7, in a volume of 20 µl. After adding 7 µg of VRR-NUC or 3.5 µg of Exonuclease VIII, the reactions were incubated in 40°C for 20 min.

S-2L exonuclease VIII is confirmed to have a strong exonuclease activity, indicated by the complete disappearance of the band corresponding to the plasmid. In contrast, VRR-NUC does not show any detectable endo- nor exonuclease behavior. It is however not that much surprising, as VRR-NUC nucleases are known to rely on a specific DNA structure with 5' overlap (*flap* structure) to exert their funtion (178).

How cyanophage S-2L selects an alien base in its DNA: a structure-function approach

Dariusz Czernecki (1,2), Pierre Legrand (3), Sandrine Rosario (1), Mustafa Tekpinar (1), Pierre Alexandre Kaminski (4) and Marc Delarue (1)*

(1) Unit of Structural Dynamics of Biological Macromolecules, UMR 3528 du CNRS, 25 rue du Docteur Roux, Institut Pasteur, 75015 Paris, France.

(2) Sorbonne Université, Collège Doctoral, ED 515, 75005 Paris, France

(3) Synchrotron SOLEIL, L'Orme des Merisiers Saint Aubin, 91192 Gif-sur-Yvette France.

(4) Unit of Biology of Pathogenic Gram Positive Bacteria, Institut Pasteur, 75015 Paris, France.

*Corresponding author.

Abstract

Bacteriophages have long been known to use modified bases in their DNA to prevent cleavage by the host's restriction endonucleases. Among them, cyanophage S-2L is unique because its genome has all its adenines (A) systematically replaced by 2-aminoadenines (Z). Here we identify a member of the PrimPol family as the sole possible polymerase and demonstrate it has both polymerase and primase activity, albeit with a slight preference for A vs Z in front of a T. Its crystal structure at 1.5 Å resolution confirms there is no structural element in the active site that could lead to the rejection of A in front of T. To resolve this contradiction, we show that a nearby gene is in fact a triphosphohydolase specific of dATP (DatZ) that leaves intact all other dNTPs, including dZTP. This explains the absence of A in S-2L genome. Crystal structures of DatZ at sub-angstrom resolution with various ligands allow to describe its mechanism as a typical two-metal-ion mechanism and to set the stage for its engineering.

Introduction

All living organisms use the same elementary bricks for their genetic material, namely four, and only four, nucleobases: adenine (A), thymine (T), guanine (G) and cytosine (C). However, certain viruses of bacteria (bacteriophages or phages) use modified bases to escape their host's defence system, especially their endonucleases [1]. Most of the observed DNA modifications concern nucleobases and occur mainly at position 5 of pyrimidines or position 7 of purines [2], which face the major groove of the DNA double helix. For instance for pyrimidines, DNA containing 5-hydroxymethylcytosine has long been known to exist in phages T2, T4 and T6 [3], along with the enzyme (deoxycytidylate hydroxymethylase) responsible for its biosynthesis [4]; more complicated post-replicative pathways of thymidine hypermodification were recently found in phages and recreated *in vitro* [5]. For purines, archaeosine, a modified 7-deaza analogue of guanine found in archaeal tRNA D-loop [6] was found in the genome of the *E. coli* siphophage 9g [7], and is possibly present in another siphophage BRET [8]; their genomes encode genes (QueC, QueD, QueE) necessary for the biosynthesis of guanine modification. Recently, three additional 7-deazaguanine analogues have been identified and characterized in the genomes of phages and archaeal viruses [9]. An important point is to distinguish between replicative and post-replicative DNA modifications: if a biosynthetic pathway can be identified for the synthesis of the triphosphate of the modified nucleotide, it is reasonable to assume that the modified base is incorporated during replication and is not the result of a post-replicative modification.

Cyanophage S-2L is a Synechococcus phage from the double-stranded DNA *Siphoviridae* family. It was first isolated and described in 1977 [10] and its genome was shown to contain no adenine nor any of its 7-deaza derivatives. Instead, it uses 2-aminoadenine (2,6-diaminopurine or Z) that has an additional amino group in position 2 compared to adenine [11]. The A:T basepair (two hydrogen bonds) is therefore replaced by the Z:T basepair that has three hydrogen bonds, as in the G:C

basepair (Fig. 1). This feature, combined with an unusually high GC content of S-2L genome, explains its exceptionally high melting point [10]. It is believed that the A-to-Z substitution arose as a form of host evasion tactics, rendering S-2L's DNA resistant to the DNA-targeting proteins of its host, especially endonucleases [12,13].

Once the S-2L's genome was sequenced, the presence of a gene homologous to an adenylosuccinate synthetase (*purA*) was noted, raising the possibility that the phage encodes in its genome the enzymes of the biosynthesis pathway of 2-aminoadenine triphosphate (dZTP; patent application EP1499713A2). An accompanying paper reports detailed structural and functional studies of *purA* orthologues (called *purZ*) in phages S-2L and ϕVC8, both containing 2-aminoadenine in their genome, and fully supports this hypothesis (Sleiman et al., submitted; see associated PDB structures of purZ in complex with its substrates). However, it remained still largely unknown how the phage S-2L incorporates the base Z in its genome, especially as no gene corresponding to a DNA polymerase could be detected. This is in contrast with the situation in the phage ϕVC8, where a DNA polymerase of the polA family has been identified [14].

Here we identify the enzyme that is responsible for genome duplication of the phage S-2L, a member of the PrimPol family, and we present its crystal structure. We confirm its primase and polymerase activities but find that the enzyme is not specific to A vs Z. Instead, we propose that the absence of A in S-2L genome is explained by a separate enzyme, an HD phosphohydrolase that specifically dephosphorylates dATP and that we named DatZ. We give a structural explanation for both the specificity and the reaction mechanism of DatZ, based on crystallographic structures determined at sub-angstrom resolution.

Results

A DNA primase-polymerase nonspecific of A or Z

Parsing the genomic sequence of cyanophage S-2L (AX955019) in the search for a protein involved in DNA replication, we identified one ORF corresponding to a member of the Archaeo-Eukaryotic Primase (AEP) superfamily, which had not been noted earlier. As discussed below, functional tests confirm both a primase and a polymerase activity for this enzyme; thus, we will refer to it as "PrimPol", similarly to its close homologues, and its gene will be referred to as "*pplA*".

AEP is the eukaryotic and archaeal counterpart of DnaG, the bacterial primase superfamily [15,16], to which it is structurally unrelated. Its members are found in all domains of life, including viruses, and are involved in several DNA transactions including not only DNA priming and replication, but also DNA repair through non-homologous end-joining (NHEJ) [16]. AEP proteins are often fused or physically interact with DNA helicases, and also with partners containing helix bundle domains (like PriCT-1, PriCT-2, PriL or PriX) that interact with the template ssDNA [15,17–20]. Particularly important for this work, it was recently shown that a phage-encoded AEP polymerase is capable of replicating the whole genome of the NrS-1 phage [21]. Although AEP is not officially included yet in the standard DNA polymerase classification encompassing polymerases from families A, B, C, D, X, Y and RT [22,23] despite an incentive to do so [16], members of the AEP family share the classical Klenow fold with families A, B and Y DNA polymerases [24].

We started by characterizing the domain organisation of PrimPol *in silico*, using DISOPRED [25]. The result indicated that the enzyme is composed of three domains, whose function was then determined individually by homology searches (Fig. 2a). The first region (1-190) corresponds to the AEP domain itself, with all crucial motifs conserved. The second region (210-300) has a strong homology with PrCT-2 domain, most probably involved in the priming activity [17]. Together they are joined by a flexible linker and form the primase-polymerase component (1-300). The C-terminal

domain (350-737) begins after another large flexible linker. BLAST searches [26] indicate it matches best the VirE family of single-stranded DNA-binding proteins of function not described in the literature [27]. However, homology detection combined with structure prediction performed with HHpred [28] finds high-scoring similarity between viral hexameric DNA helicase structures, the closest being from bovine papillomavirus (2GXA).

We found no other detectable DNA polymerase in the S-2L's genome and went on to assay the DNA polymerase activity of PrimPol. Specifically, we looked for its ability to selectively incorporate the base Z in front of an instructing base T, discarding the dATP present in the host cell's dNTP pool and avoiding the A:T basepair altogether. We cloned and overexpressed the synthetic gene of PrimPol in *E. coli* and tested its polymerase activity *in vitro*. To study the specificity towards A and Z, we used dsDNA with a dT_{12} oligomer as the 5' overhang of the template strand and either dATP or dZTP in the reactional mixture. We tested a range of different conditions, varying temperature, pH, DNA, nucleotide and enzyme concentrations, as well as various divalent ions (Fig. 2b-d) that are usual cofactors in DNA and RNA polymerases [29]. All assays indicate that S-2L PrimPol is capable of incorporating both nucleotides in front of T, accepting A more readily than Z. We also noted that the presence of Mn^{2+} ions induces limited terminal transferase activity, as observed for some other DNA polymerases such as the human pol μ from the pol X family [30]; for another, more distantly-related AEP, this activity was observed even with Mg^{2+} ions [31]. We also overexpressed truncated versions of the enzyme, PP-N300 and PP-N190, corresponding to the primase-polymerase core and polymerase domain, respectively. We did not observe any change in the catalytic activity for PP-N300 (Suppl. Fig. S1a), confirming the necessary and sufficient role of the core domain during DNA synthesis. In another test, we showed that PP-N300 can synthesise *in vitro* the first 124 nucleotides of its own native gene, with both ATGC and ZTGC mixtures (Suppl. Fig. S1b).

Finally, we confirmed that PrimPol exhibits DNA-dependent DNA primase activity, dependent on the length and specific sequence of the template, but independently of the third domain of the enzyme (Suppl. Fig. S2).

Structural analysis of the AEP domain of S-2L PrimPol

Using BLAST, we identified 129 other sequences with high similarity to the AEP domain of PrimPol (PP-N190). We aligned them and visualised the conservation status of crucial residues and motifs described in previous reports (Fig. 3a); their function is described further below.

We could crystallize PP-N190 and solve its structure at 1.5 Å resolution (PDB ID: 6ZP9; Suppl. Table 1, Fig. 3b), using phase information from SeMet derivative crystals. Ca^{2+} ions were mandatory in the mother liquor to obtain crystals. As expected, the protein has a classical AEP fold. All crucial residues cluster together in the catalytic site of the domain (Fig. 3c). Y63, E85, D87, T112, K115, H118, D146, R157 are conserved across all AEPs (or have biochemically similar counterparts), and their function is well established in the superfamily. Residue Y63 plays the role of a steric gate for ribonucleotides, allowing only dNTPs in the catalytic site [32]. Residues E85, D87 (that can vary to Asp and Glu, respectively) coordinate a divalent metal ion (M^{2+}) in the B site, that positions the triphosphate of the incoming nucleotide (dNTP) during polymerisation; this triphosphate is further stabilized by interactions with T112, K115, H118 and R157 (possibly varying respectively to Ser, Arg, Asn and Lys) [33–36]. Residue D146 along with residues E85, D87 and the dNTP's α-phosphate coordinate another M^{2+} ion in the A site, making it possible to add the incoming dNTP to the primer strand of the nascent nucleic acid through the two-metal-ion mechanism [33,37,38]. The three negatively charged residues E85, D87 and D146 are crucial for the polymerase and primase activity, as shown in the related human PrimPol [39]. Importantly, in S-2L PP-N190 we noticed a significant positional shift of residue D87 compared to other AEP structures, along with the conservation among the close relatives of the

neighbouring residue D88, which is exposed to the solvent. We propose that in the case of PP-N190 and its relatives the residue D87 plays a direct role in the primase activity together with D88. Instead, its usual role in binding M^{2+} ion in the B site could be fulfilled by S116; we investigate this possibility in detail further below. Finally, although residue H163 lies further apart from the triphosphate, its high conservation and covariance with positions R157 and H118 was noticed in a recent study [17]. In human PriS, the mutation of the corresponding residue (H324) to alanine partially inhibited the enzymatic activity, a result that was explained by the presence of a water molecule that links it to the triphosphate [34].

Due to the presence of divalent calcium ions in all crystallisation conditions, we could not soak the crystals with nucleotides which immediately precipitate; transferring crystals to a solution devoid of Ca^{2+} dissolved the crystals in a matter of seconds. On the other hand, there are several AEP structures with bound ligands available in the PDB, including DNA and (d)NTPs. Based on the three structures with DNA (3H25, 3PKY, 5L2X), the nucleic acid apparently bends in an L-shape over the open catalytic site (Suppl. Fig. S3a). Additionally, the incoming (d)NTP's conformation is largely conserved across all 8 unique AEP structures with a bound nucleotide (PDB IDs: 1V34, 2ATZ, 2FAQ, 3PKY, 5L2X, 5OF3, 6JON, 6R5D). In all cases, the catalytic site is open to the solvent and there is no selection on the incoming nucleotides; after superposition with these structures, PP-N190 presents no structural feature that could lead to a Z vs A specificity during the polymerase reaction.

A possible mechanism for the primase catalytic site

We next tried to understand how the PrimPol works in the primase mode. Relying on structure of human PrimPol [35], we could build a model of S-2L PrimPol AEP domain with a Mg^{2+} ion placed in the classical site B in the presence of two nucleotide triphosphates in the elongation (polymerase) and initiation (primase) sites, as well as with another Mg^{2+} ion in an additional metal binding site (C) between residues D87

and D88 (Suppl. Fig. S3b). Using this initial model, we conducted molecular dynamics (MD) simulations to investigate the stability of the complex in the catalytic site. We observed during these simulations that the side-chain of S116 was coordinating the Mg^{2+} ion in the B site, along with the (predicted by homology) residue E85 (Suppl. Fig. S3c). Strictly conserved between closely related PP-N190 relatives but not across the AEP superfamily, S116 can apparently replace the shifted D87 residue, rather than contacting the γ-phosphate of the incoming nucleotide as seen for its counterpart in human PrimPol [35]. Additionally, the Mg^{2+} ion placed at site C between residues D87 and D88 was stable during the 212 ns-long MD simulation, and interacts with the γ-phosphate of the nucleotide in the initiation site. The possible change of D88 to Asn or to His observed in related AEP domains retains the capacity of divalent metal ion binding and further supports the functional nature of this position. We propose that this additional ion binding site "C" is important in the positioning and charge neutralisation of the 5' nucleotide during primase activity of PrimPol involving two nucleotide triphosphates.

In conclusion, while the discovery of PrimPol encoded in S-2L's genome explains how the phage could replicate its genome, functional and structural studies show it cannot discriminate A against Z. Therefore, it remains to be explained how Z gets incorporated in the genome of S-2L instead of A.

DatZ: a triphosphohydrolase specific of dATP

We subsequently revisited other genes susceptible to intervene during the phage genome replication. We found that one ORF in the immediate vicinity of *purZ* encodes a 175 aa protein belonging to the HD-domain phosphohydrolase family [40]. Enzymes from this family are known to dephosphorylate standard deoxynucleotide monophosphates (dNMPs) and can also act as a triphosphatase on dNTPs, as well as some close nucleotide analogues [41,42]. After purification of the S-2L HD phosphohydrolase overexpressed in *E. coli*, we tested its activity by pre-incubating it

with the reactional mixture for the aforementioned DNA polymerization assay, before adding PrimPol. We observed that the presence of the phosphohydrolase prevented polymerization with dATP, but did not affect the polymerisation with dZTP (Fig. 2d).

We interpreted this behaviour as the result of a specific dATP triphosphohydrolase activity, therefore suggesting to call the enzyme DatZ. We confirmed this hypothesis by incubating DatZ with different nucleotide triphosphates and analysing the reaction products by HPLC analysis (Fig. 4). dATP was rapidly degraded into dA; however, under the same conditions there was no dephosphorylation of ATP, dZTP, nor of all other standard dNTPs (dGTP, dTTP or dCTP). We also found no phosphorylase activity on dADP or dAMP substrates (Suppl. Fig. S4a). Marginal tri-dephosphorylation products of dZTP start to appear only after a prolonged incubation (75x longer than for dATP) or in excess of DatZ concentration. Contrary to OxsA phosphohydrolase [42], we did not observe a sequential dephosphorylation, but a one-step reaction directly from dNTPs to dNs, never detecting any intermediate phosphorylation states in the course of the reaction.

Our finding that S-2L DatZ is a specific dATP triphosphohydrolase offers a simple explanation of how the phage avoids incorporating adenine in its genome.

DatZ structure at 0.86 Å resolution: general description

Using X-ray crystallography, we determined three structures of S-2L DatZ with its substrate, the reaction product and its metal cofactors, the second one at sub-angstrom resolution. They constitute the first structures of a viral HD phosphohydrolase, and the third HD phosphohydrolase to be described in atomic details, after *E. coli* YfbR [43] and *B. megaterium* OxsA [42].

First, we present a 0.86 Å resolution structure of S-2L DatZ bound to dA, the product of dephosphorylation of dATP in solution (PDB ID: 6ZPA; Suppl. Table 1). The electron

density allowed to build the whole protein as well as 219 water molecules around the DatZ chain (175 aa), which is roughly the number expected for this resolution limit [44]. Although several hydrogens are discernible at such a resolution, the usual limit for their experimental allocation is 0.8 Å [45]; they were therefore refined using a riding model. Each monomer of DatZ takes a globular form composed predominantly of α- and 3_{10}-helices (70% and 4% respectively), with no β-strands (Fig. 5a). The base moiety of dA snugly fits in the catalytic pocket below a relatively flexible element (as indicated by higher B-factors), with the P79 residue on its tip (Fig. 5b). A catalytic divalent ion is found in the vicinity of dA's free 5'-OH group, even though no divalent ion was added in buffers during purification or crystallisation. In the catalytic site, the side chain of residue I22 is ideally positioned to sterically exclude the amino group in position 2 of the purine ring of G or Z and provides an immediate explanation for the observed specificity of the enzyme. In addition, W20 side chain constitutes a steric hindrance for the 2' hydroxyl group of any ribose-based nucleotide.

Concerning the oligomeric state of DatZ, we found that *in crystallo* it arranges in a compact toroidal hexamer with a D3 symmetry, where neighbouring subunits are flipped (Fig. 5c). Such a shape emerges from two partially hydrophobic, self-interacting protein sides (A:A and B:B), with a large surface of interaction – 1358.6 $Å^2$ and 959.0 $Å^2$. We confirmed the hexameric stoichiometry of DatZ *in vitro* with complementary techniques, *i.e.* DLS and analytical ultracentrifugation leading to 5.9 (± 0.1) monomers per oligomer, assuming a perfectly globular shape. The whole hexamer is particularly rigid, as judged from the overall very low B-factors (Fig. 5d), consistent with the ultrahigh diffraction limit for DatZ crystals.

A two-metal-ion mechanism of DatZ

In the literature, there is some uncertainty as to which divalent cation plays a catalytic role in HD phosphohydrolases: the structure of OsxA suggested the presence of one fixed Co^{2+} ion coordinated by the protein and one transient Mg^{2+} interacting

with the triphosphate [42]. The YfbR enzyme was shown to be active with Co^{2+} and less with Mn^{2+}, Cu^{2+} and Zn^{2+} [41], while OxsA is roughly equally active with Co^{2+}, Co^{2+}/Mg^{2+} and Mn^{2+}, but not Zn^{2+} [42].

In S-2L DatZ, the first detected metal ion, occupying the site "A" in the 0.86 Å resolution structure is coordinated by residues H34, H66, D67 and D119; two water molecules, also present in the Co^{2+}-bound structure (see below), complete a typical octahedral coordination shell and fit well the electron density map. Both the position and coordination of ion A^{2+} are identical to what is observed in other known HD phosphohydrolase structures. An excitation x-ray energy scan and an anomalous double-difference Fourier map analysis consistently point to the presence of a Zn^{2+} ion in this site. Its coordination geometry is less common than the usual tetragonal one, but not atypical [46]. The fact that no additional divalent ions were added during protein purification indicates a high affinity of DatZ to Zn^{2+}. Zn^{2+} is present in *E. coli* grown on LB medium [47] at a level comparable to the one found in vivo in cyanobacteria [48].

We then solved a second structure of DatZ co-crystallised with dATP and 10 mM $CoCl_2$ (PDB ID: 6ZPB; Suppl. Table 1, Suppl. Fig. S5a) and noticed the presence of a second, still undescribed metal ion binding site, that we call "B". This site is not the one observed in OxsA structure, although it lies in the vicinity of the first site (5.2 Å apart) as well. Both Co^{2+} ions are coordinated octahedrally: in site A, the binding geometry is the same as described above for Zn^{2+}, while in site B the coordination is mediated by residues E70, D75, the O5' of dA and three water molecules. The presence of the two Co^{2+} ions was confirmed by a strong anomalous signal in the corresponding Fourier difference map.

Finally, we solved a third structure of DatZ, this time with dATP bound (PDB ID: 6ZPC; Suppl. Table 1, Suppl. Fig. S5b) but no divalent ion(s), obtained by adding EDTA to the enzyme before crystallizing it with dATP. In this structure, we could observe the residues K81 and K116 neutralising the negative charge of β- and γ-phosphates. We

still find a Zn^{2+} ion in the A-site as shown by its anomalous signal, although not fully occupied and only penta-coordinated. We assume that this change in coordination, intermediate between tetrahedral and octahedral and also commonly observed for Zn^{2+} [46], is the result of the presence of a triphosphate.

Superposing all new structures with both co-factors (divalent ions) and the substrate allows to propose a complete catalytic mechanism of DatZ (Fig. 6). Similarly to alkaline phosphatase and 3'-5' exonuclease [49], DatZ uses a typical two-metal-ion mechanism to dephosphorylate dATP. While the ion B^{2+} stabilizes leaving O5' atom and one oxygen of the α-phosphate (P_α), ion A^{2+} positions a hydroxide (OH^-) in an attacking position opposite to O5'. Then, by interacting with OH^- the α-phosphate passes through a penta-coordinate intermediate, forming an unstable oxyanion stabilized by the R19 residue. Finally, the bond O5'-P_α is broken and a new one, P_α-OH, is created.

We checked by HPLC that DatZ is active in a buffer containing Mg^{2+} as the sole added divalent metal ion and we observed that the enzyme stays active *in crystallo* with no additional divalent ions. Two additional crystal structures showed that Zn^{2+} in site A is replaced by Co^{2+} in excess of the latter (20 mM $CoCl_2$), but is retained in elevated Mg^{2+} concentrations (50 mM $MgSO_4$), as determined through anomalous signal analysis at the corresponding wavelengths (data not shown).

The active site of DatZ: conservation and mutagenesis

A number of phages that contain a close homologue of *purZ* gene in their genome (Sleiman et al., submitted) also contain a homologue of *datZ*. Looking for the conservation of residues crucial for both a dATPase activity and absence of dZTPase activity, as identified by the present structural studies, we built a multialignment of those closely-related DatZ sequences (Suppl. Fig. S6). We found that all residues stabilizing both catalytic metal ions are strictly conserved, as well as R19, K81 and

K116 interacting with α-, β- and γ-phosphates. Residues W20, I22 and P79, interacting with the base, are conserved or involve conservative substitutions. Additionally, residues Q29, A32 and G74 are strictly conserved among close DatZ homologues, highlighting their possible importance for protein structure (ternary or quaternary) and/or its dynamics.

With the intention of engineering a dNTPase with a selectivity shifted towards dZTP, we cloned, expressed and tested DatZ I22A mutant, designed to make room for the additional amino group of Z in the binding pocket. We observed a significant relaxation of its specificity (Fig. 2d, Suppl. Fig. S4b). The mutant's dATPase activity is clearly reduced and still does not show any intermediate product. The additional space created for the 2-amino group of dZTP has the desired effect of raising the dZTPase activity to the point of becoming detectable, albeit still very low. The dGTPase activity remains undetectable, indicating that the selectivity towards an amino group in position 6 of the purine ring is maintained.

Discussion

The immediate neighbours of PrimPol in the S-2L genome are also replication-related proteins (exonuclease VIII, Snf2 helicase and VRR-NUC endonuclease), and all have a high level of sequence identity with Mediterranean uvMED phages' corresponding proteins. In contrast, those viruses contain neither *purZ* nor *datZ* genes – they share with S-2L only their replicative machinery, and not the additional apparatus that enables the A-to-Z switch. Interestingly, S-2L PrimPol is also related to cyanobacterial enzymes: notably, sequence motifs in the AEP polymerase core correspond perfectly to those of All3500-like family [17], with almost all of the high-scoring matches coming from cyanobacteria genus. Such a finding supports the idea that *pplA*, the gene of PrimPol, may have been exchanged between cyanophages and their hosts.

Due to the divergent nature of the AEP superfamily, its classification is far from trivial. The universal presence of its members, encompassing all three domains of life, viruses and plasmids, testifies about its ancient origin [17]. Advanced sequence-based computational methods divided the superfamily into four clades [15]: AEP proper, NCLDV-herpesvirus primase, PrimPol, and BT4734-like. In another approach using sequence clustering [17], AEPs were distributed into multiple groups, with the newly defined PrimPol-PV1 supergroup. S-2L PrimPol belongs to the *Anabaena* (*Nostocaceae*) All3500-like family within the PrimPol clade or the PrimPol-PV1 supergroup, depending on the classification.

A search with PrimPol in the Dali server [50] identified all structures of AEP available in the PDB. However, due to excessive divergence of the superfamily, the structure-based multialignment approach, applied below for DatZ, was not reliable. Instead, we adopted the geometry-based analysis proposed by Dali. Both the dendrogram and the non-hierarchical clustering method (Supp. Fig. S7) distinguish two major, well-defined groups: archaeo-eukaryotic replicative PriS primases and bacterial NHEJ primases (LigC/D), belonging to the *AEP proper clade* defined previously [15]. The remaining set contains PrimPols with more distant homology. The strongest link between S-2L PrimPol and any other member of the AEP family is with the plasmidic RepB' (3H20), highlighting the connection between All3500-like and RepB' clusters within the PrimPol-PV1 supergroup [17]. Additionally, the previously undescribed campylobacterial AEP represented by HP0184 from *H. pylori* (2ATZ) is systematically placed together with them, hinting that they may share a common ancestor.

In general, in spite of the modest set size of 15 unique AEP structures, PrimPols are clearly much more widespread and diverse than the PriS and NHEJ primases which have more specific roles. Our preliminary analysis suggests that the ancestor of S-2L PrimPol was acquired from its cyanobacterial host.

Concerning DatZ, we performed a multialignment of all available HD phosphohydrolase structures with PROMALS3D (Supp. Fig. S8), thus avoiding purely sequence-based errors. There is a strict conservation of all residues binding metal ion A across all representatives, along with metal B-binding E70 residue and R19 that stabilizes the reaction intermediate. There are two singular cases where the D75 B-site binding residue can change to E or H, but chemically both are capable of metal ion coordination. Prominently, the human HD phosphohydrolase HDDC2 (HD domain-containing protein 2) shows a metal coordination identical to the one seen in S-2L DatZ; it is the only other homologue structure with two ions (Mg^{2+}) present in both sites A and B (PDB ID 4DMB). Although it was hypothesised that during the nucleophilic attack this glutamic acid would act as a proton donor through a water bridge [43], here we provide evidence that it participates in metal B binding instead. Interestingly, the E72A YfbR mutant (corresponding to DatZ E70) was described as having lost its phosphohydrolase activity. Lastly, the residue E93 is almost completely structurally conserved, with the only exception of OxsA, and its position along the sequence is shifted one α-helix turn in DatZ; it remained undetected by previous sequence alignments with close viral DatZ homologues probably due to an intrinsic low precision in this region without structural support. E93 places its sidechain in the catalytic pocket, but too far away to interact directly with the phosphate γ or the divalent metal ion B^{2+} (6.5 and 7.8 Å, respectively). We suggest that this glutamic acid may instead facilitate the free phosphates' trafficking between the catalytic pocket and the solvent.

Using the multialignment data, we constructed a structurally-informed phylogenetic tree of HD phosphohydrolases (Suppl. Fig. S9). Aside from following the typical distribution into the tree domains of life, it suggests that the ancestor of DatZ was acquired from a bacterial variant; the closest DatZ homologs found in BLAST represent the phyla of γ-proteobacteria and firmicutes (excluding the immediate viral clade), in conformity with this hypothesis.

Although diverse in sequence, the monomeric structures of the other known HD phosphohydrolases are very similar (Supp. Fig. S10a), with an average RMSD on C_α atoms of 2.75 Å. Despite the fact that only a dimer was described for related bacterial HD phosphohydrolases [42,43], we discovered that the same hexameric quaternary state could be found by generating their symmetry-related mates using the space-group symmetry operators (Supp. Fig. S10b). In fact, a high multimeric state (>3) has been also reported *in vitro* for YfbR [41], compatible with our hypothesis.

As all residues crucial for the reaction in DatZ are conserved or replaced by similar residues in other structures, we suggest that the two-metal-ion mechanism described above is universal for all HD phosphohydrolases, completing previous reports by the identification of metal ion site B and correcting the role of residue E70 counterparts (Suppl. Fig. S10c). Interestingly, OxsA replaced the positively charged K116 with E129 bearing negative charge; we propose that it is this exception that forces OxsA to accommodate a third divalent metal ion, unobserved in DatZ, to efficiently neutralise the charge of the triphosphate.

In conclusion, we note that the strategy adopted by the phage S-2L phage is most probably shared with related phages containing homologous *datZ* and *purZ* genes. It is very similar to the strategy adopted by the T2, T4 and T6 phages that contain a substantial amount of hydroxymethylcytosine, relying on a dCTP triphosphatase to also shift the pool of available dNTPs in their host cell [4].

In the future, it will be interesting to see if the genes uncovered here and in related work (Sleiman et al., submitted) are sufficient for transferring 2-aminoadenine to the genomes of other organisms.

Acknowledgements

We thank the Crystallogenesis and Crystallography Platform (PF6) of Institut Pasteur for help in crystallisation, and the Molecular Biophysics Platform (PFBMI) for protein quality control, ultracentrifugation and DLS experiments. We thank M. Hollenstein and his group for the use of their HPLC system. We also thank staff from PROXIMA-1 and PROXIMA-2 beamlines for help in data collection and SOLEIL (Saint-Aubin, France) and for provision of synchrotron radiation facilities. MT thanks Programme Pause from Collège de France for financial support. We thank P. Marlière for getting us interested in the S-2L phage in the first place as well as for numerous discussions, and Valérie Pezo for pointing out to us other phosphohydrolases of the dUTPase family in related phages.

Author contributions

MD directed the study. DC and MD designed the study. DC, PL, SR and MT performed experiments. DC, PL, SR, MT, PAK and MD analysed the data. DC and MD wrote the manuscript.

Competing Interests

The authors declare they have no competing financial interests.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Bibliography

1. Gommers-Ampt, J. H. *et al.* β-d-glucosyl-hydroxymethyluracil: A novel modified base present in the DNA of the parasitic protozoan T. brucei. *Cell* **75**, 1129–1136 (1993).
2. Iyer, L. M., Zhang, D., Maxwell Burroughs, A. & Aravind, L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res* **41**, 7635–7655 (2013).
3. Wyatt, G. R. & Cohen, S. S. The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine. *Biochemical Journal* **55**, 774–782 (1953).
4. Koerner, J. F., Smith, M. S. & Buchanan, J. M. Deoxycytidine Triphosphatase, an Enzyme Induced by Bacteriophage Infection. *J. Biol. Chem.* **235**, 2691–2697 (1960).
5. Lee, Y.-J. *et al.* Identification and biosynthesis of thymidine hypermodifications in the genomic DNA of widespread bacterial viruses. *PNAS* **115**, E3116–E3125 (2018).
6. Gupta, R. Halobacterium volcanii tRNAs. Identification of 41 tRNAs covering all amino acids, and the sequences of 33 class I tRNAs. *J. Biol. Chem.* **259**, 9461–9471 (1984).
7. Kulikov, E. E. *et al.* Genomic Sequencing and Biological Characteristics of a Novel Escherichia Coli Bacteriophage 9g, a Putative Representative of a New Siphoviridae Genus. *Viruses* **6**, 5077–5092 (2014).
8. Ngazoa-Kakou, S. *et al.* Complete Genome Sequence of Escherichia coli Siphophage BRET. *Microbiol Resour Announc* **8**, e01644-18 (2019).
9. Hutinet, G. *et al.* 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nat Commun* **10**, 1–12 (2019).
10. Kirnos, M. D., Khudyakov, I. Y., Alexandrushkina, N. I. & Vanyushin, B. F. 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **270**, 369 (1977).
11. Santhosh, C. & Mishra, P. C. Electronic spectra of 2-aminopurine and 2,6-diaminopurine: phototautomerism and fluorescence reabsorption. *Spectrochimica Acta Part A: Molecular Spectroscopy* **47**, 1685–1693 (1991).

12. Szekeres, M. & Matveyev, A. V. Cleavage and sequence recognition of 2,6-diaminopurine-containing DNA by site-specific endonucleases. *FEBS Letters* **222**, 89–94 (1987).
13. Bailly, C. & Waring, M. J. The use of diaminopurine to investigate structural properties of nucleic acids and molecular recognition between ligands and DNA. *Nucleic Acids Res* **26**, 4309–4314 (1998).
14. Solís-Sánchez, A. *et al.* Genetic characterization of ØVC8 lytic phage for Vibrio cholerae O1. *Virology Journal* **13**, 47 (2016).
15. Iyer, L. M., Koonin, E. V., Leipe, D. D. & Aravind, L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33**, 3875–3896 (2005).
16. Guilliam, T. A., Keen, B. A., Brissett, N. C. & Doherty, A. J. Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Research* **43**, 6651–6664 (2015).
17. Kazlauskas, D. *et al.* Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. *Journal of Molecular Biology* **430**, 737–750 (2018).
18. Geibel, S., Banchenko, S., Engel, M., Lanka, E. & Saenger, W. Structure and function of primase RepB' encoded by broad-host-range plasmid RSF1010 that replicates exclusively in leading-strand mode. *PNAS* **106**, 7810–7815 (2009).
19. Liu, B. *et al.* A primase subunit essential for efficient primer synthesis by an archaeal eukaryotic-type primase. *Nat Commun* **6**, 1–11 (2015).
20. Yan, J., Holzer, S., Pellegrini, L. & Bell, S. D. An archaeal primase functions as a nanoscale caliper to define primer length. *PNAS* **115**, 6697–6702 (2018).
21. Zhu, B. *et al.* Deep-sea vent phage DNA polymerase specifically initiates DNA synthesis in the absence of primers. *PNAS* **114**, E2310–E2318 (2017).
22. Braithwaite, D. K. & Ito, J. Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res* **21**, 787–802 (1993).

23. Raia, P., Delarue, M. & Sauguet, L. An updated structural classification of replicative DNA polymerases. *Biochemical Society Transactions* **47**, 239–249 (2019).
24. Mönttinen, H. A. M., Ravantti, J. J. & Poranen, M. M. Common Structural Core of Three-Dozen Residues Reveals Intersuperfamily Relationships. *Mol Biol Evol* **33**, 1697–1710 (2016).
25. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
27. Citovsky, V., Vos, G. D. & Zambryski, P. Single-Stranded DNA Binding Protein Encoded by the virE Locus of Agrobacterium tumefaciens. *Science* **240**, 501–504 (1988).
28. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237–2243 (2018).
29. Steitz, T. A., Smerdon, S. J., Jager, J. & Joyce, C. M. A unified polymerase mechanism for nonhomologous DNA and RNA polymerases. *Science* **266**, 2022–2025 (1994).
30. Domínguez, O. *et al.* DNA polymerase mu (Pol μ), homologous to TdT, could act as a DNA mutator in eukaryotic cells. *The EMBO Journal* **19**, 1731–1742 (2000).
31. Gill, S. *et al.* A highly divergent archaeo-eukaryotic primase from the Thermococcus nautilus plasmid, pTN2. *Nucleic Acids Res* **42**, 3707–3719 (2014).
32. Díaz-Talavera, A. *et al.* A cancer-associated point mutation disables the steric gate of human PrimPol. *Sci Rep* **9**, 1–13 (2019).
33. Zhu, H. *et al.* Atomic structure and nonhomologous end-joining function of the polymerase component of bacterial DNA ligase D. *PNAS* **103**, 1711–1716 (2006).
34. Kilkenny, M. L., Longo, M. A., Perera, R. L. & Pellegrini, L. Structures of human primase reveal design of nucleotide elongation site and mode of Pol α tethering. *PNAS* **110**, 15961–15966 (2013).

35. Rechkoblit, O. *et al.* Structure and mechanism of human PrimPol, a DNA polymerase with primase activity. *Science Advances* **2**, e1601317 (2016).
36. Guo, H. *et al.* Crystal structures of phage NrS-1 N300-dNTPs-Mg2+ complex provide molecular mechanisms for substrate specificity. *Biochemical and Biophysical Research Communications* **515**, 551–557 (2019).
37. Brissett, N. C. *et al.* Structure of a Preternary Complex Involving a Prokaryotic NHEJ DNA Polymerase. *Molecular Cell* **41**, 221–231 (2011).
38. Holzer, S. *et al.* Structural Basis for Inhibition of Human Primase by Arabinofuranosyl Nucleoside Analogues Fludarabine and Vidarabine. *ACS Chem. Biol.* **14**, 1904–1912 (2019).
39. Calvo, P. A. *et al.* The invariant glutamate of human PrimPol DxE motif is critical for its Mn2+-dependent distinctive activities. *DNA Repair* **77**, 65–75 (2019).
40. Aravind, L. & Koonin, E. V. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends in Biochemical Sciences* **23**, 469–472 (1998).
41. Proudfoot, M. *et al.* General Enzymatic Screens Identify Three New Nucleotidases in Escherichia coli BIOCHEMICAL CHARACTERIZATION OF SurE, YfbR, AND YjjG. *J. Biol. Chem.* **279**, 54687–54694 (2004).
42. Bridwell-Rabb, J., Kang, G., Zhong, A., Liu, H. & Drennan, C. L. An HD domain phosphohydrolase active site tailored for oxetanocin-A biosynthesis. *PNAS* **113**, 13750–13755 (2016).
43. Zimmerman, M. D., Proudfoot, M., Yakunin, A. & Minor, W. Structural Insight into the Mechanism of Substrate Specificity and Catalytic Activity of an HD-Domain Phosphohydrolase: The 5′-Deoxyribonucleotidase YfbR from Escherichia coli. *Journal of Molecular Biology* **378**, 215–226 (2008).
44. Nittinger, E., Schneider, N., Lange, G. & Rarey, M. Evidence of Water Molecules—A Statistical Evaluation of Water Molecules Based on Electron Density. *J. Chem. Inf. Model.* **55**, 771–783 (2015).

45. Woińska, M., Grabowsky, S., Dominiak, P. M., Woźniak, K. & Jayatilaka, D. Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science Advances* **2**, e1600192 (2016).
46. Dokmanić, I., Šikić, M. & Tomić, S. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. *Acta Crystallographica Section D* **64**, 257–263 (2008).
47. Outten, C. E. & O'Halloran, and T. V. Femtomolar Sensitivity of Metalloregulatory Proteins Controlling Zinc Homeostasis. *Science* **292**, 2488–2492 (2001).
48. Rajeshwari, K. & Rajashekhar, M. Biochemical Composition of Seven Species of Cyanobacteria Isolated from Different Aquatic Habitats of Western Ghats, Southern India. *Brazilian Archives of Biology and Technology* **54**, 849–857 (2011).
49. Kim, E. E. & Wyckoff, H. W. Reaction mechanism of alkaline phosphatase based on crystal structures: Two-metal ion catalysis. *Journal of Molecular Biology* **218**, 449–464 (1991).
50. Holm, L. Benchmarking fold detection by DaliLite v.5. *Bioinformatics* doi:10.1093/bioinformatics/btz536.
51. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**, 28–33 (2003).
52. Sauguet, L., Raia, P., Henneke, G. & Delarue, M. Shared active site architecture between archaeal PolD and multi-subunit RNA polymerases revealed by X-ray crystallography. *Nature Communications* **7**, 12227 (2016).
53. Weber, P. *et al.* High-Throughput Crystallization Pipeline at the Crystallography Core Facility of the Institut Pasteur. *Molecules* **24**, 4451 (2019).
54. Kabsch, W. XDS. *Acta Cryst D* **66**, 125–132 (2010).
55. Legrand, P. *XDS Made Easier (2017) GitHub repository*.
56. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst D* **75**, 861–877 (2019).

57. Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Cryst D* **66**, 479–485 (2010).
58. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* **31**, 671–690 (2010).
59. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **34**, 2135–2145 (2013).
60. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).
61. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802 (2005).
62. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47**, W636–W641 (2019).
63. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).
64. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **14**, 1188–1190 (2004).
65. Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295–2300 (2008).
66. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **42**, W320–W324 (2014).
67. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).
68. *The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.*

Materials and methods

Identification of genes of interest

The genomic sequence of cyanophage S-2L was obtained from NCBI's database (AX955019). Potential ORFs were identified using ORFFinder [51] (>150 nt, genetic code 11). Targeted ORFs were assessed for known homologous proteins using BLAST. Protein disorder of AEP was predicted with DISOPRED [25].

Protein expression and purification

Synthetic genes for expressed proteins were optimized for *E. coli* and synthesized using ThermoFisher's GeneArt service. Genes were cloned into modified RSF1-Duet expression vector with a TEV-cleavable N-terminal 14-histidine tag [52] using New England Biolabs and Anza (Thermo Fisher Scientific) enzymes. Shorter versions of PrimPol (PP300 and PP190) were obtained by adding overhangs with codon STOP and corresponding cleavage site through standard PCR with designed oligonucleotides (Eurogentec); mutagenesis of DatZ was done using designed oligonucleotides and QuikChange II Site-Directed Mutagenesis Kit (Agilent). *E. coli* BL21-CodonPlus (DE3)-RIPL cells (Agilent) were separately transformed with engineered plasmids. Bacteria were cultivated at 37°C in LB medium with appropriate antibiotic selection (kanamycin and chloramphenicol), and induced at OD=0.6-1.0 with 0.5 mM IPTG. After incubation overnight at 20°C, cells were harvested and homogenized in suspension buffer: 50 mM Tris-HCl pH 8, 400 mM NaCl, 5 mM imidazole. After sonication and centrifugation of bacterial debris, proteins of interest were isolated from corresponding lysate supernatants by purification on Ni-NTA column (same buffer with 500 mM imidazole). They were further diluted to 150 mM NaCl and repurified on HiTrap Heparin (for PrimPol) or HiTrap Q (for DatZ) columns (1 mM NaCl in elution buffer). Histidine tags were removed from the proteins by incubation with his-tagged TEV enzyme overnight. After removing TEV on Ni-NTA column, proteins were further purified on Superdex 200 10/300 column with 25 mM Tris-HCl pH 8, 150 mM NaCl (for PrimPol-N190 crystallisation a 16 mM concentration of NaCl was used).

All purification columns were from Life Sciences. Protein purity was assessed on an SDS gel (BioRad). The enzymes were concentrated to 7-19 mg ml^{-1} with Amicon Ultra 10k and 30k MWCO centrifugal filters (Merck) and stored directly at -80°C, with no glycerol added.

DNA polymerase and primase assays

The polymerase activity tests, if not stated otherwise for a particular condition, were executed in 20 mM Tris-HCl pH 7 and 5 mM $MgCl_2$, with 3 µM of FAM-marked dT_{12} or $dT_{10}GG$ DNA template, 1.5 µM of 15 nt DNA primer complementary to template upstream sequence, 500 µM dNTP mix, 0.5 µM of AEP (10 min of incubation) and 1 µM of DatZ (8 min) at 37°C. The Klenow polymerase used as a control was at 5 U in 50 µl (10 min incubation). Polymerase gene replication test was conducted similarly, with 3 µM of template and primer labelled radioactively on 5' end [α-32P]; PP-300 at 42.3 µM and Klenow polymerase at 4 U in 20 µl were incubated for 15 min. The primase activity test was conducted in 20 mM Tris-HCl pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 5 mM $MgSO_4$, 5 mM $MgCl_2$ and 0.1% Triton X-100, with 10 µM of DNA template, 250 µM of dNTP mix (ATGC), 25 µM of dATP [α-32P] and 2 µM of PrimPol (15 min of incubation) at 50°C. Before adding protein, DNA was hybridized by heating up to 95°C and gradually cooling to reaction temperature. Reactions were blocked with the addition of a buffer containing 10 mM EDTA, 98% formamide, 0.1% xylene cyanol and 0.1% bromophenol blue, and stored in 4°C. Products were preheated at 95°C for 10 min, before being separated with polyacrylamide gel electrophoresis and visualised by FAM fluorescence or radioactivity on Typhoon FLA 9000 imager. All oligonucleotides were ordered from Eurogentec, chemicals from Sigma-Aldrich, Klenow polymerase from Takara Bio, standard dNTPs from Fermentas (Thermo Fisher Scientific) and dZTP from TriLink BioTechnologies.

Nucleotide HPLC analysis

1 µM of DatZ or its mutant was incubated at 37°C for 10 min with 500 µM of the respective dNTP, in a buffer containing 20 mM Tris pH 7 and 5 mM $MgCl_2$. Reaction

products were separated from the protein using 10 000 MWCO Vivaspin-500 centrifugal concentrators and stored in -20°C. Products and standards were assayed separately, using around 40 nmol of each for anion-exchange HPLC on DNA-PAC100 (4x50 mm) column (Thermo Fisher Scientific). After equilibration with 150 µl of a suspension buffer (25 mM Tris-HCl pH 8, 0.5% acetonitrile), nucleotides were injected on the column and eluted with 3 min of isocratic flow of the suspension buffer followed by a linear gradient of 0-200 mM NH_4Cl over 10 min (1ml min^{-1}). Eluted nucleotides were detected by absorbance at 260 nm. High-purity nucleotides and chemicals were bought from Sigma Aldrich, and HPLC-quality acetonitrile was from Serva.

Crystallography and structural analysis

All crystallization conditions were screened using the sitting drop technique on an automated crystallography platform [53] and were reproduced manually using the hanging drop method with ratios of protein to well solution ranging from 1:2 to 2:1. PrimPol-N190 was screened at 14.5 mg ml^{-1} in 4°C. Elongated rods grew over 2 days in 100 mM $CaCl_2$, 20% w/v PEG 8k (40%) and 5% v/v isopropanol (100%) buffered with 100 mM MES pH 6. DatZ was screened at 12-17 mg ml^{-1} with a molar excess of 1.2 of dATP at 18°C. Big, symmetric crystals grew rapidly over 1-2 days in 1.5 M $LiSO_4$ buffered with 100 mM HEPES pH 7.5. All crystals were soaked in a solution containing 70% crystallization buffer and 30% glycerol and frozen in liquid nitrogen. Crystallographic data was collected at the SOLEIL synchrotron in France (beamlines PROXIMA-1 and PROXIMA-2), processed by XDS [54] with the XDSME [55] pipeline and refined in Phenix [56]. The structure of PrimPol-N190 was solved by SAD technique using SeMet derivative of the protein and data sets collected at the selenium edge (0.980655 Å) using the SHELX C/D/E programs [57]. The structure of DatZ was solved by the sulphur-SAD (S-SAD) technique at 1.771203 Å wavelength. One Zn^{2+} ion was unambiguously identified by its peak in the anomalous double difference Fourier map at 40 sigmas with datasets collected at 9.67 and 9.66 keV (Zn peak and pre-edge).

DatZ ultrahigh resolution structure was obtained by merging 3 individual datasets taken on the same crystal. Structures of DatZ with bound Co^{2+} and dATP were obtained by growing crystals with 10 mM $CoCl_2$ and 10 mM EDTA, respectively (the latter at pH 7). Anomalous signal analysis for Zn^{2+} ion retention in presence of Co^{2+} or Mg^{2+} was performed (unpublished data) at the respective K-edge wavelengths of 1.5981 Å and 0.7749 Å, for Co and Zn respectively.

Molecular dynamics simulations of PP-N190

Force field parameters of dCTP were obtained using CGenFF [58]. The parameter penalty and the charge penalty were zero, indicating that the parameters can be used safely without any modification. CHARMM36 parameter set was used for the rest of the system [59]. Topologies of the structures were prepared with psfgen module of VMD [60]. After the topology construction, the structures were solvated in a triclinic box with a distance of at least 11 Å to the box edges and TIP3P solvent model. The systems were neutralized with Na^+ and Cl^- ions, and the ion concentration was set to 0.15 M. Then, a 50000 step conjugate gradient minimization procedure was carried out. The minimized systems were heated up to 300 K with 0.001 K steps. An NPT equilibration procedure followed the heating. The equilibration time was 2 ns and the time step was 2 fs. The equilibration temperature (300 K) was controlled with Langevin thermostat and the pressure (1 atm) was controlled with Langevin barostat. Production runs were 212 ns long, with the remaining parameters of production runs identical to the equilibration stage parameters. All of the molecular dynamics simulations were performed with NAMD version 2.13 [61].

Sequence and structure alignments, phylogeny

Close relatives of *pplA* and *datZ* were identified by BLAST searches, and aligned with Clustal Omega [62] (PrimPol) or the default MUSCLE algorithm in MEGA X software [63] (DatZ); ; sequence logos were made with WebLogo [64]. Structures homologous to PrimPol and DatZ available in PDB were identified using Dali server [50]; Dali was further used for pairwise RMSDs determination and geometry analysis. The

tendencies observed for AEP superfamily clustering were maintained whether the analysis involved whole structures or only AEP cores, and whether the dataset was complete or not. The sequences of DatZ and other structures from HD phosphohydrolase family were aligned in PROMALS3D [65] using structural data supplemented by full protein sequences, excluding not-superimposable N- and C-termini. Multialignment images were prepared with ESPript 3 [66]. Maximum-likelihood phylogenetic tree of HD phosphohydrolases based on their structural multialignment was prepared in MEGA X with default parameters, taking 100 bootstrap replications. All protein structures were visualised with Chimera [67] and Pymol [68].

Figures

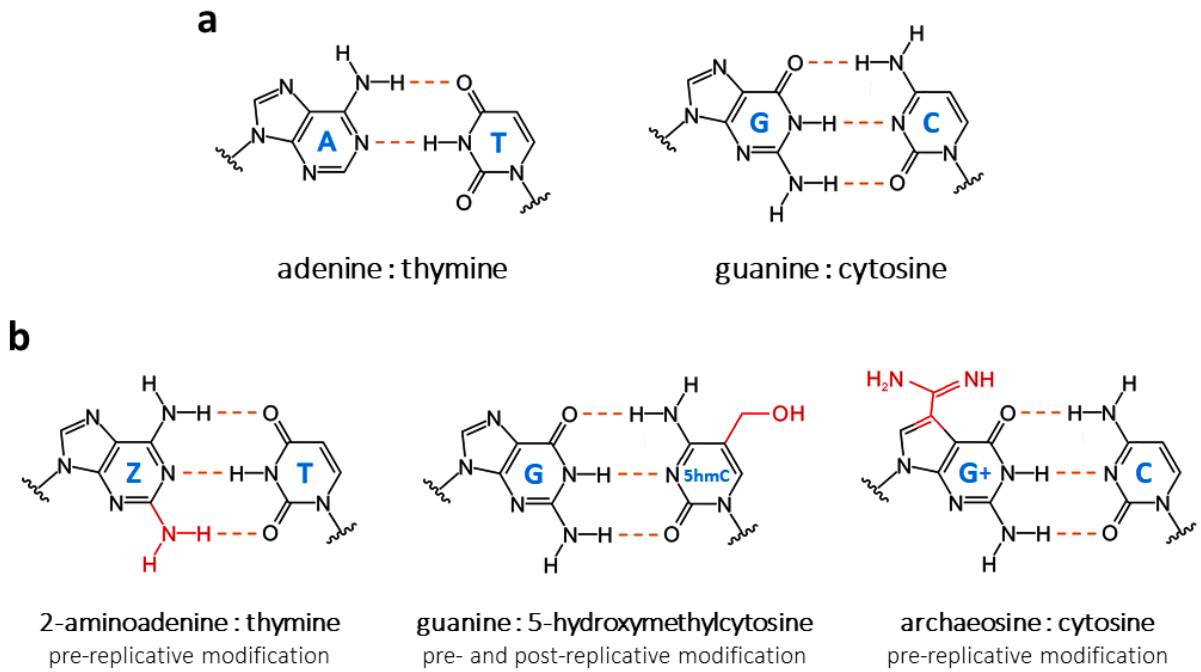


Figure 1. **Watson-Crick base pairs and natural variations thereof. a.** Classical DNA base pairs, universal to all three domains of life and most viruses. **b.** Other types of base pairs with three hydrogen bonds found in some organisms and viruses. Additional chemical groups are in red. 2-aminoadenine:thymine (Z:T, left); guanine:5-hydroxymethylcytosine (G:5hmc, center); archaeosine:cytosine (G+:C, right). The Z:T pair, first found in cyanophage S-2L, replaces completely the usual A:T pair in its genome.

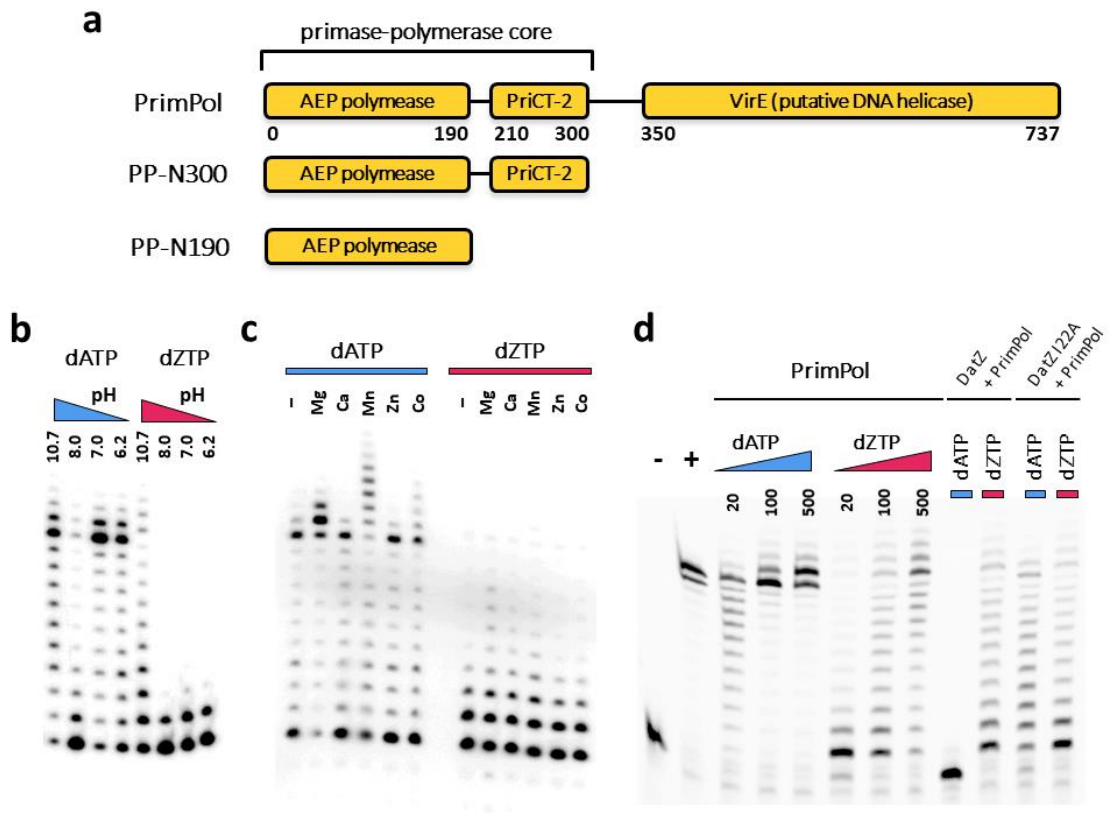


Figure 2. **Functional characterisation of S-2L PrimPol. a.** Schematic diagram of S-2L PrimPol constructs showing its different domains with their respective amino-acid range (to scale). **b-d.** Results of DNA polymerase activity tests of S-2L PrimPol with either dATP (blue) or dZTP (magenta) as the incoming dNTP, using dT_{12} (b and d) or $dT_{10}GG$ (c) overhang templates. **b.** Different buffers with various pHs. **c.** Effect of different divalent ions. **d.** Effect of growing concentrations of nucleotides (lanes 3-8) and pre-incubation of reactional mixture for DatZ WT (lanes 9-10) and I22A mutant (lanes 11-12). Lanes 1-2 represent 1) a negative control, without any polymerase, and 2) a positive control, with *E. coli* Pol I (Klenow fragment).

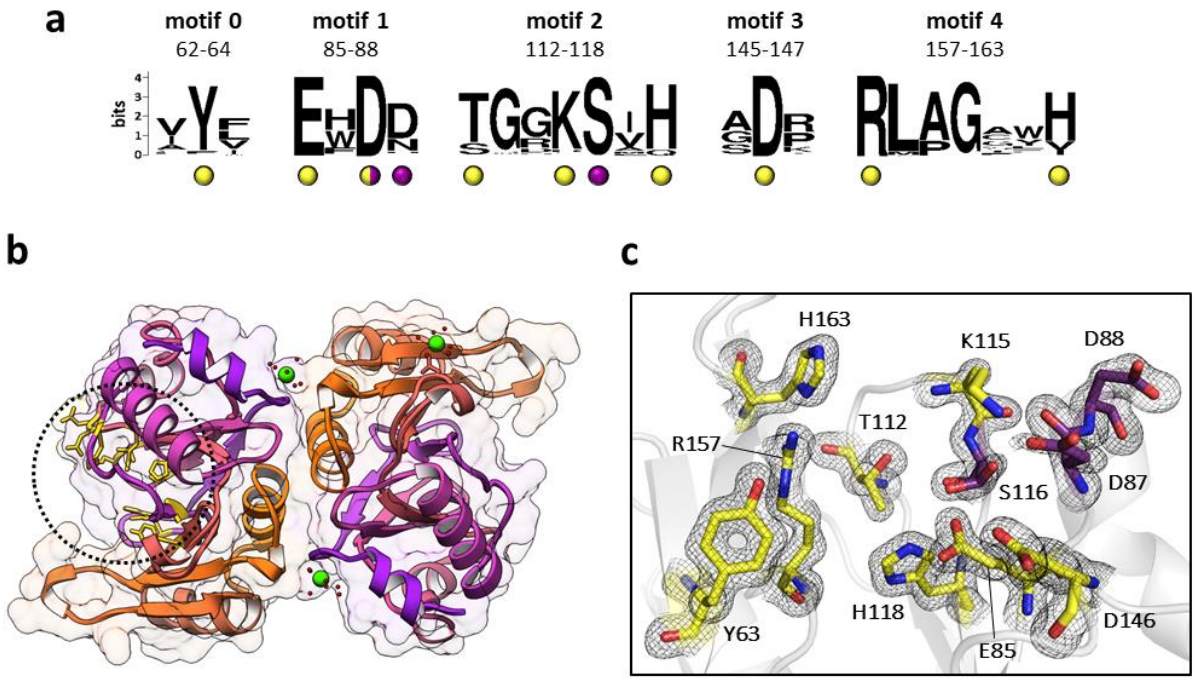


Figure 3. **AEP domain of S-2L PrimPol: conserved residues and their structural context. a.** Five AEP motifs of PP-N190 close homologues. In addition to previous motif classifications [35,17], the steric gate tyrosine is included as motif 0, and motifs 1 and 2 are extended. Numbers on top of the sequence blocks indicate their amino acid range according to S-2L PrimPol. Highlighted are conserved residues with the function described for other AEPs in the literature (yellow dot underneath), and residues yet undescribed with proposed function (purple dot). The double-hatted residue D87 would be involved in both polymerase (known) and primase (suggested) activities. **b.** Structure of PP-N190 in ribbon and surface representation, with two symmetric molecules in the crystallographic asymmetric unit, each coloured with an orange-purple gradient. Calcium ions are shown by green spheres, with water molecules forming their hydration shells shown as red ones. The catalytic site of molecule A is shown in yellow stick representation and indicated with a dotted circle. **c.** Zoom on the catalytic site of PP-N190. Residues highlighted in (a) are shown in stick representation and labelled, maintaining the same colour code. The experimental electron density around those residues (black mesh) is contoured at 1 sigma.

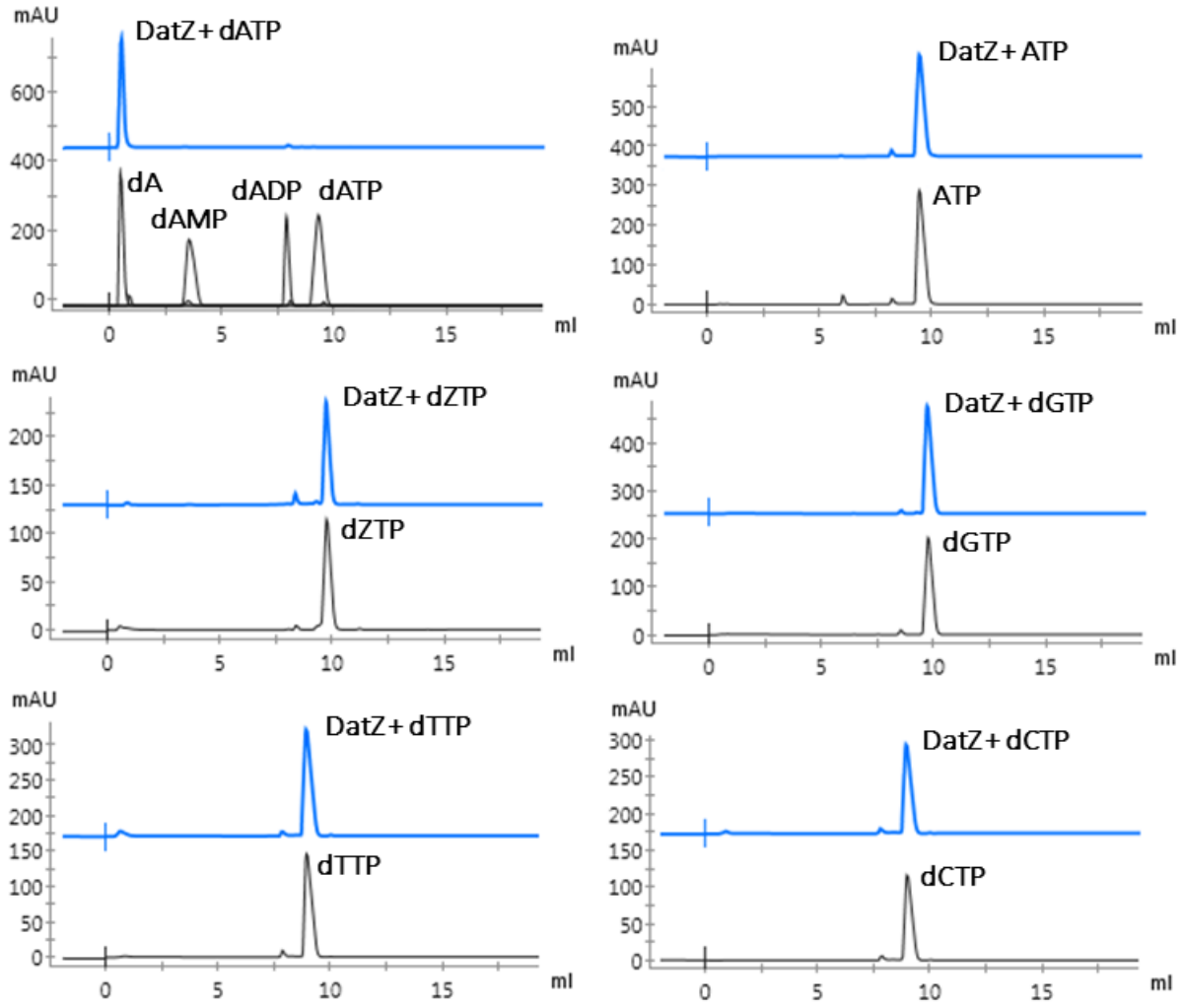


Figure 4. **HPLC analysis of S-2L DatZ dephosphorylation products.** Nucleotide standards are in black, products eluted after incubation of the corresponding triphosphates with DatZ are in blue. Each sample was eluted separately, using an amount of 40 nmol. The enzyme is active exclusively with dATP and removes from it all phosphates: it is therefore a triphosphohydrolase specific of dATP, or dATPase.

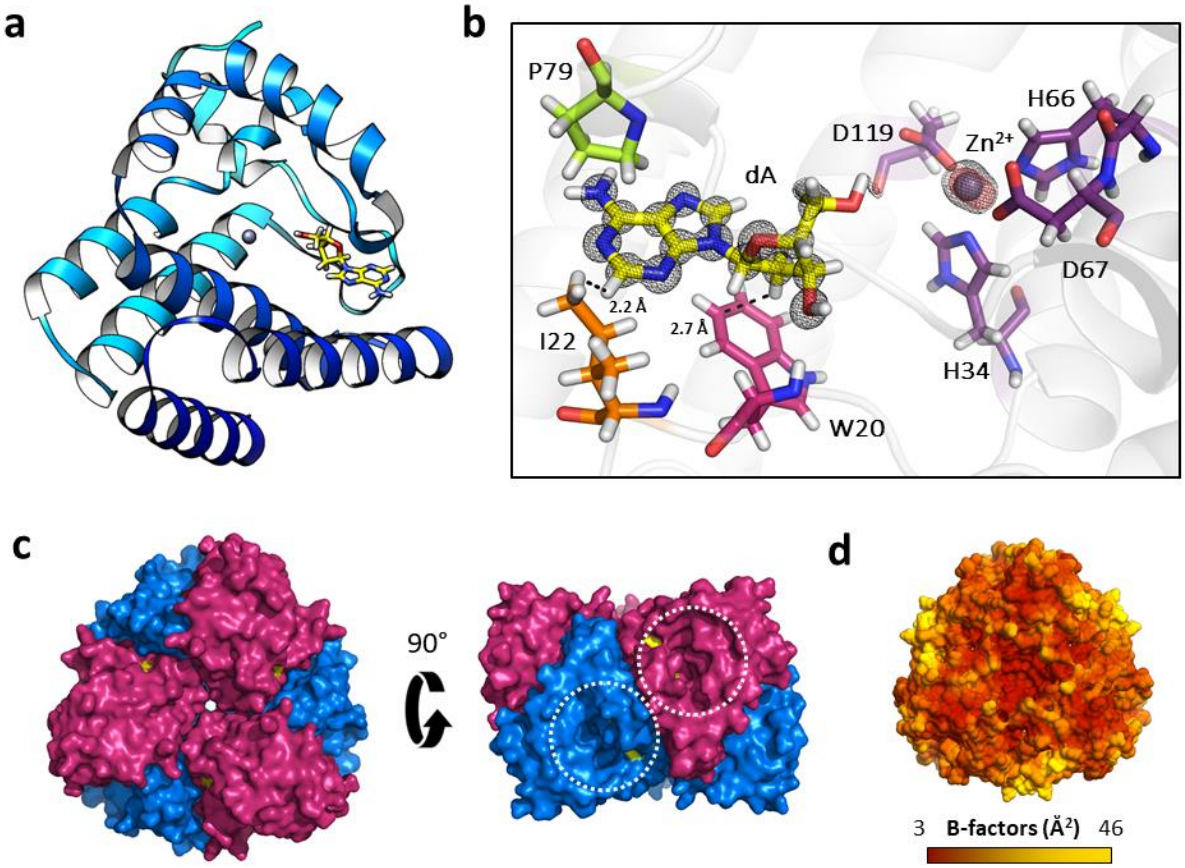


Figure 5. **Three-dimensional structure of S-2L DatZ. a.** Ribbon representation of a DatZ monomer in a light blue-dark blue gradient, with bound dA in stick (yellow). The Zn^{2+} ion is shown as a grey sphere. **b.** A close-up on the catalytic pocket of DatZ with the experimental electron density contoured at 2.5 sigmas around bound ligands: dA and Zn^{2+} (black mesh). Additionally, the anomalous density at Zn^{2+} absorption edge (red mesh) is contoured at 10 sigmas. Residue I22 (orange) provides direct specificity towards the adenine nucleobase, creating a steric hindrance for chemical groups in position 2 of the purine ring. Other residues highlighted in the text are Zn^{2+}-coordinating ones (purple), W20 (magenta) and P79 (lime). **c.** Structure of the full DatZ hexamer, top and side views, in surface representation. Blue and purple monomers form a compact, particularly stable disc in an alternating, zigzagging pattern. Two of the six symmetrical cavities leading to buried dA molecules (yellow) are visible in the side view and highlighted by the white dotted circles. **d.** Surface representation of DatZ hexamer coloured by the experimental B-factors (dark red-yellow gradient, hydrogen atoms omitted), with the scale bar below. Highest temperature factors correspond to the two helices above dA.

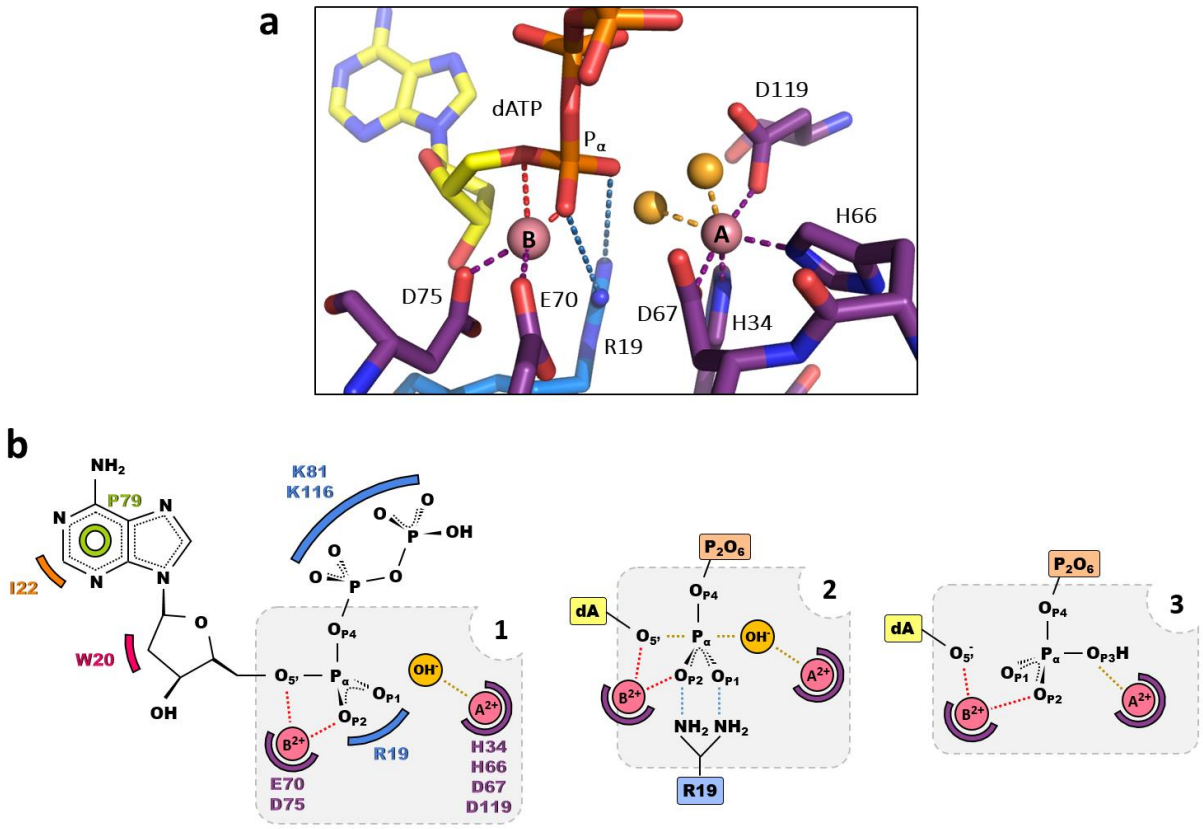
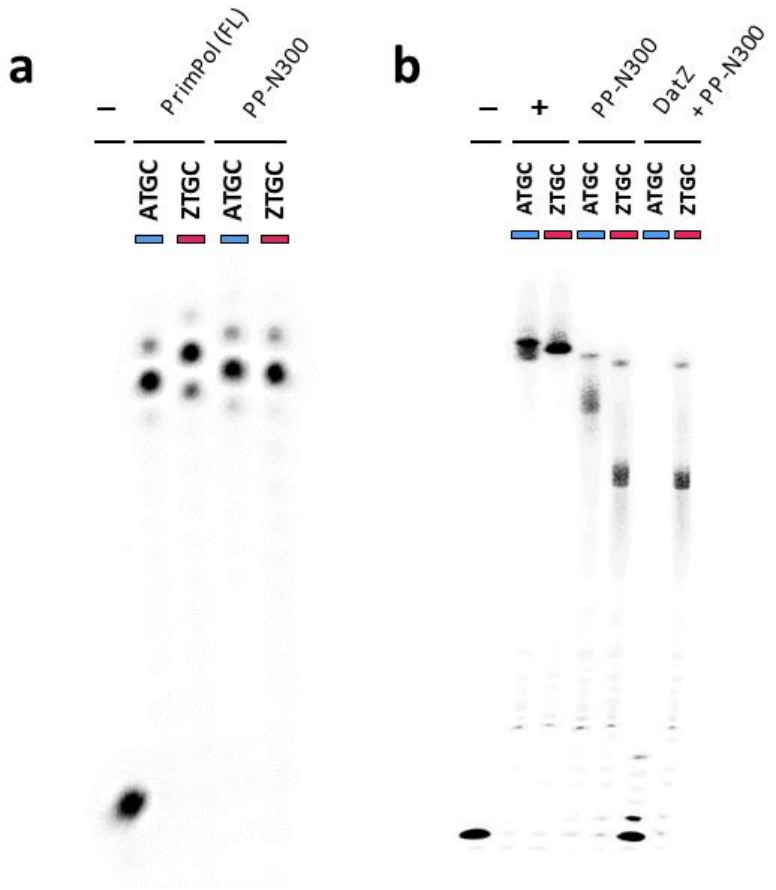
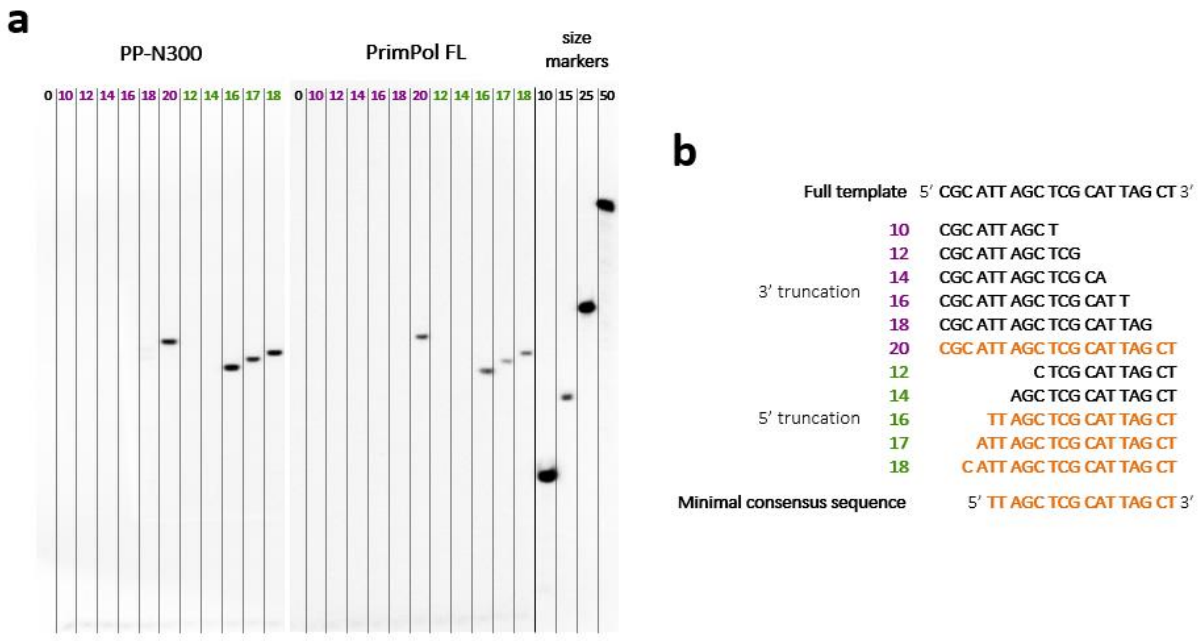


Figure 6. **Catalytic centre of S-2L DatZ with the substrate and cofactors and the mechanism of tri-dephosphorylation. a.** Model of the reaction centre made by superposition of two of the structures solved in this work. The first structure defines dATP (in yellow) and residue R19 interacting with the α-phosphate (blue); hydrogen atoms were omitted for consistency. The second structure provides catalytic ions A and B (magenta spheres), bound water molecules that are likely to take part in the reaction (gold) and the metal coordinating residues (purple). Interacting atoms, ions and groups of interests are shown by dashed lines of corresponding colour. The distance between the two Co^{2+} ions is 5.2 Å. **b.** Schematic diagram of DatZ reaction under two-metal-ion mechanism with the initial substrates (1), intermediate (2) and products (3). The colour code is identical as in (a), extended by base-stabilising P79 (lime), sugar-specificity-conferring W20 (magenta), 2-amino-specificity-conferring I22 (orange) and triphosphate-neutralizing K81 and K116 (blue) residues. In this diagram, a hydroxide ion (OH^-) is proposed for the nucleophile.

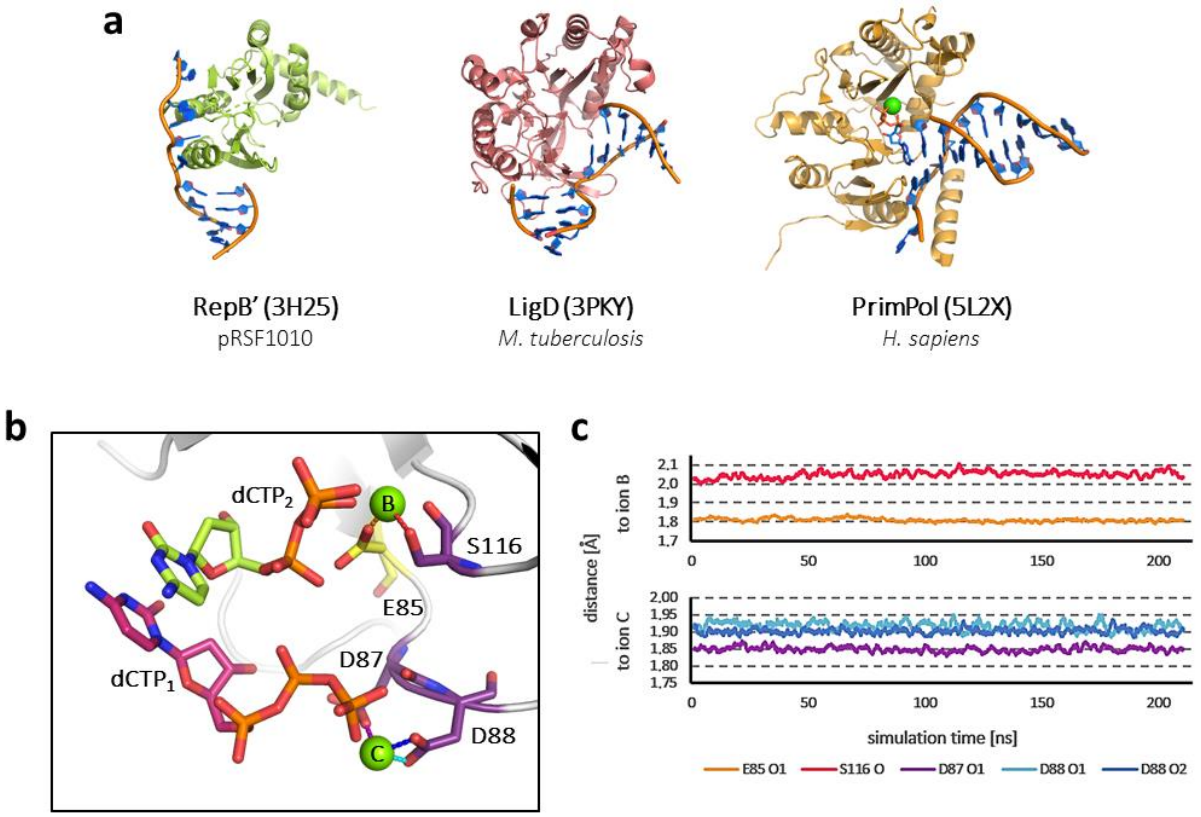
Supplementary Data



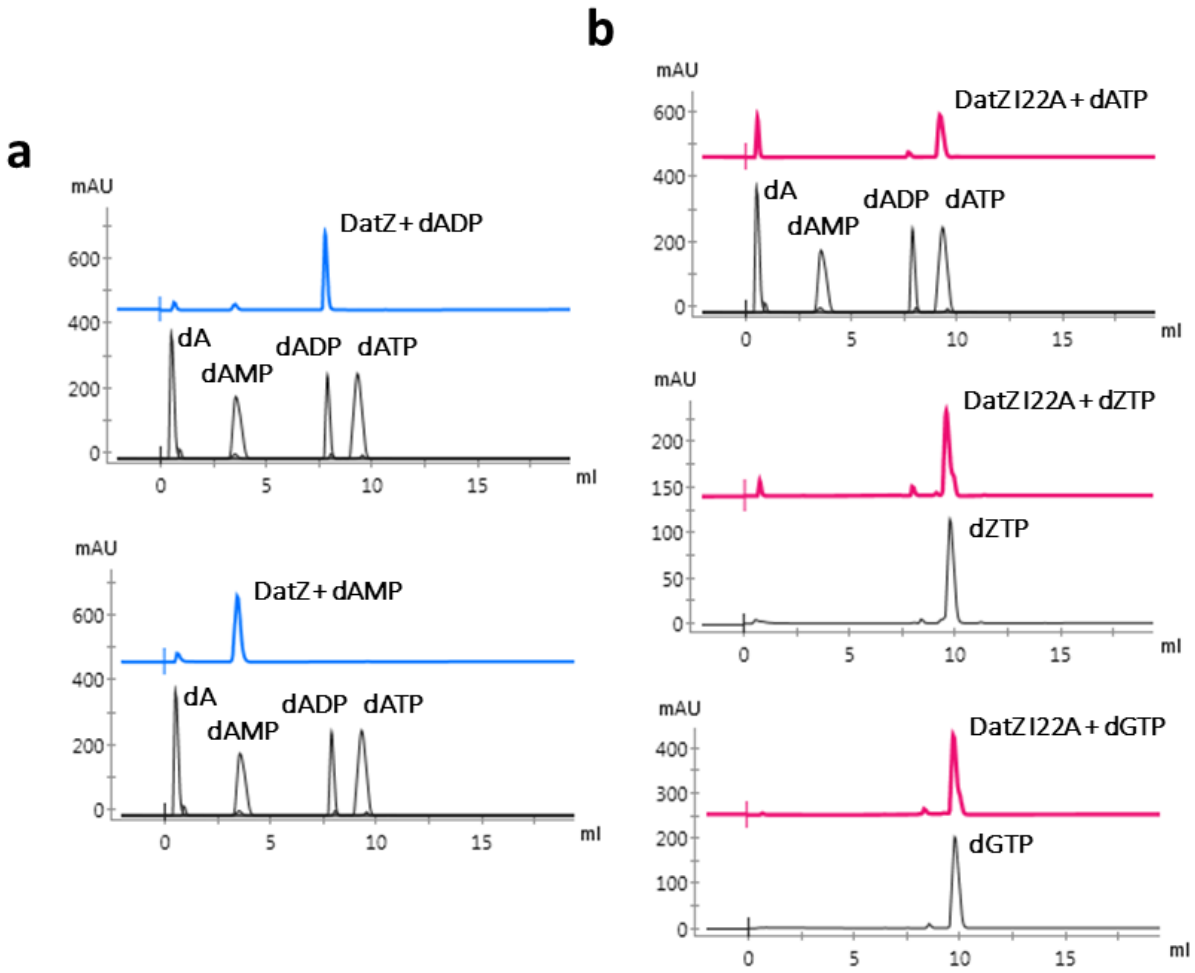
Supplementary Figure S1. Additional catalytic tests on S-2L PrimPol constructs, using ATGC (blue) or ZTCG (magenta) nucleotide mixes. a. Activity tests of the full-length PrimPol (FL) and the truncated one (PP-N300), with a negative control without any polymerase in the first lane. The incubation time was 20 min and the concentrations were 50 nM for the $dT_{10}GG$ 5'-overhang template, 250 µM of each dNTP and 1 µM of PrimPol. **b.** Polymerisation assay for the first 124 nt of PrimPol's native gene. Results are shown for a negative control without any polymerase (lane 1), a positive control with *E. coli* Pol I (Klenow fragment) (lanes 2-3), PP-N300 polymerase without (lanes 4-5), or with pre-incubation of the reactional mixture with DatZ (lanes 6-7). The polymerisation step was allowed to proceed for 15 min with 42.3 µM of PP-N300.



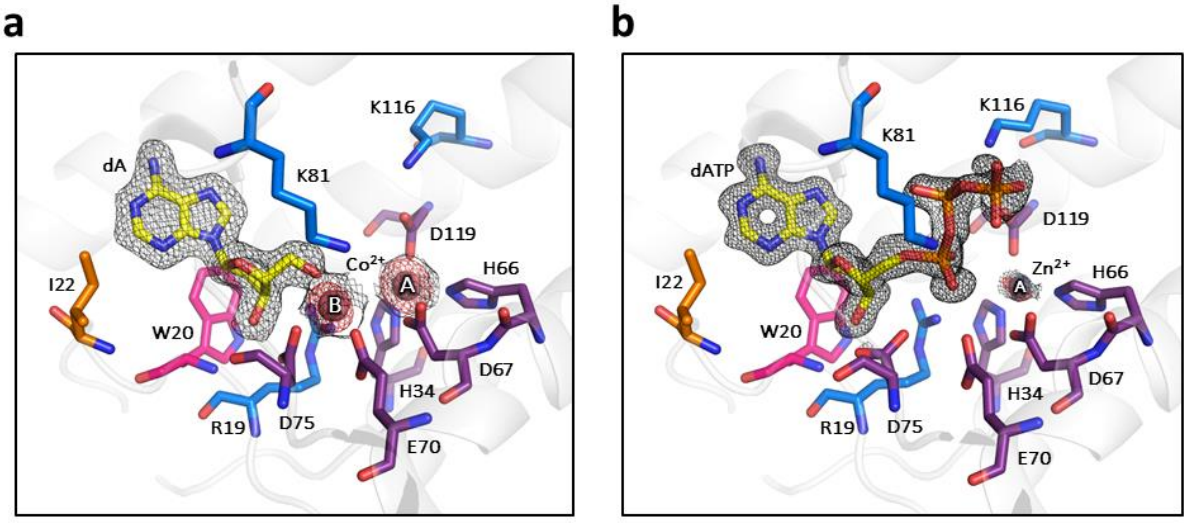
Supplementary Figure S2. DNA-dependent DNA primase activity of S-2L PrimPol. a. Primase activity tests of two constructs: full-length PrimPol (FL) and the truncated one (PP-N300) on a range of DNA templates, with negative controls without the template in the first lanes and size markers in the last four. Lengths of the templates, truncated from 3' (purple) and 5' ends (green) are shown above the corresponding lanes. The presence or absence of the last domain of PrimPol does not change its priming activity. **b.** Detailed sequences of templates from the previous panel, using the same colour code. In orange are sequences undergoing priming with S-2L PrimPol. The primase activity depends on the template length and its sequence: the minimal consensus sequence that could be identified is shown in the bottom line.



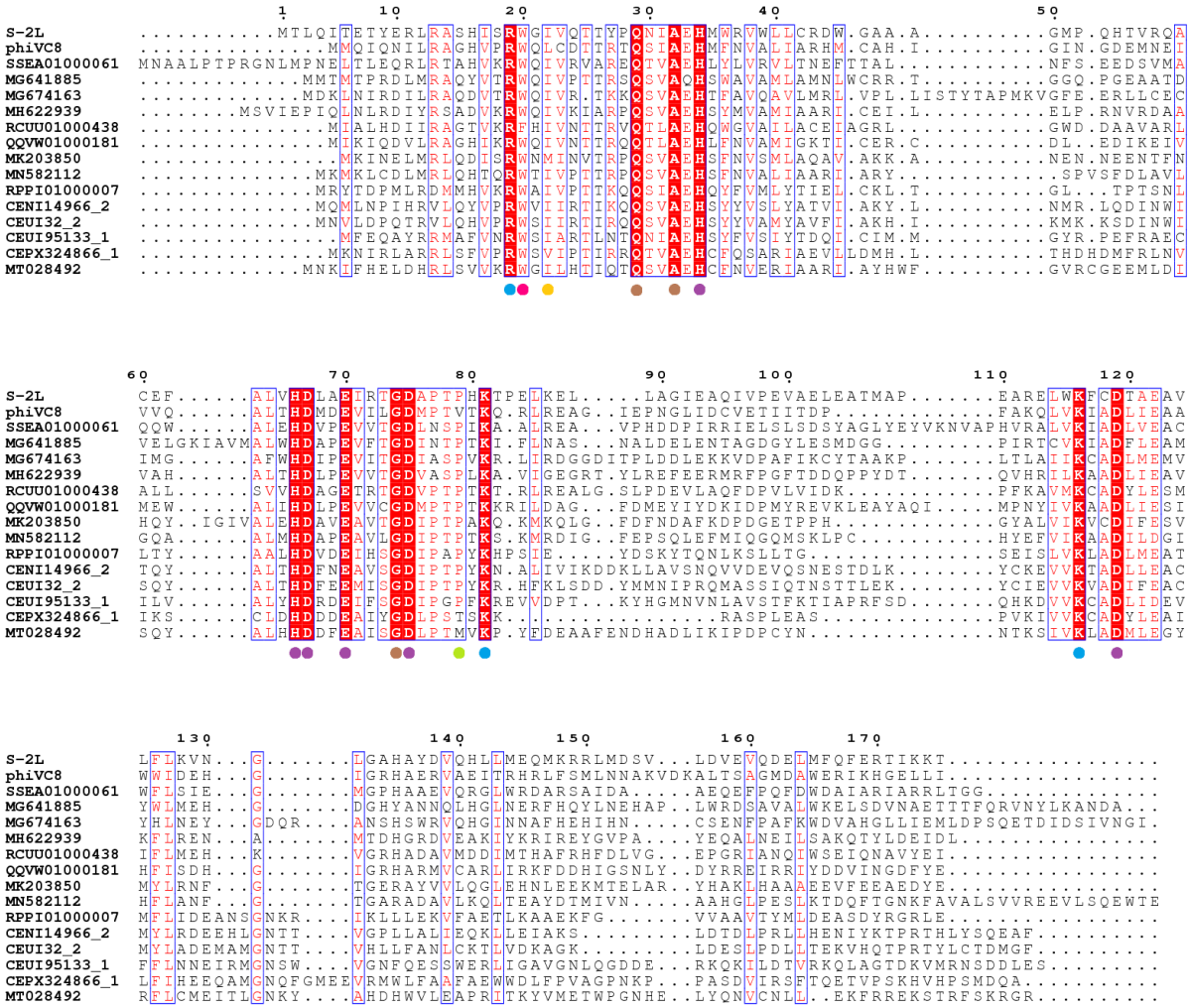
Supplementary Figure S3. AEP and its ligands. a. Three AEP structures with bound dsDNA, viewed from the same perspective. Below are: the protein name, PDB code of the structure and the organism or plasmid of origin. The DNA molecule seems to bend in an L-shape at the catalytic site. **b.** A model of PP-N190 with two Mg^{2+} ions (green) bound in B and C sites and two nucleotides in the initiation (magenta) and elongation (lime) sites, obtained after energy minimization step. Residues interacting with Mg^{2+} ions in a way previously undescribed are in purple, with ionic bonds visualised by the coloured dashed lines. The novel ion in site C interacts with γ-phosphate of the nucleotide in the initiation site. **c.** Distances between the residues shown in (b) and bound ions, using the same bond colour code. They were measured in the course of 212 ns of the exemplary simulation (two ions) and averaged in the 2 ns frame. The binding is stable and similar across all simulations.



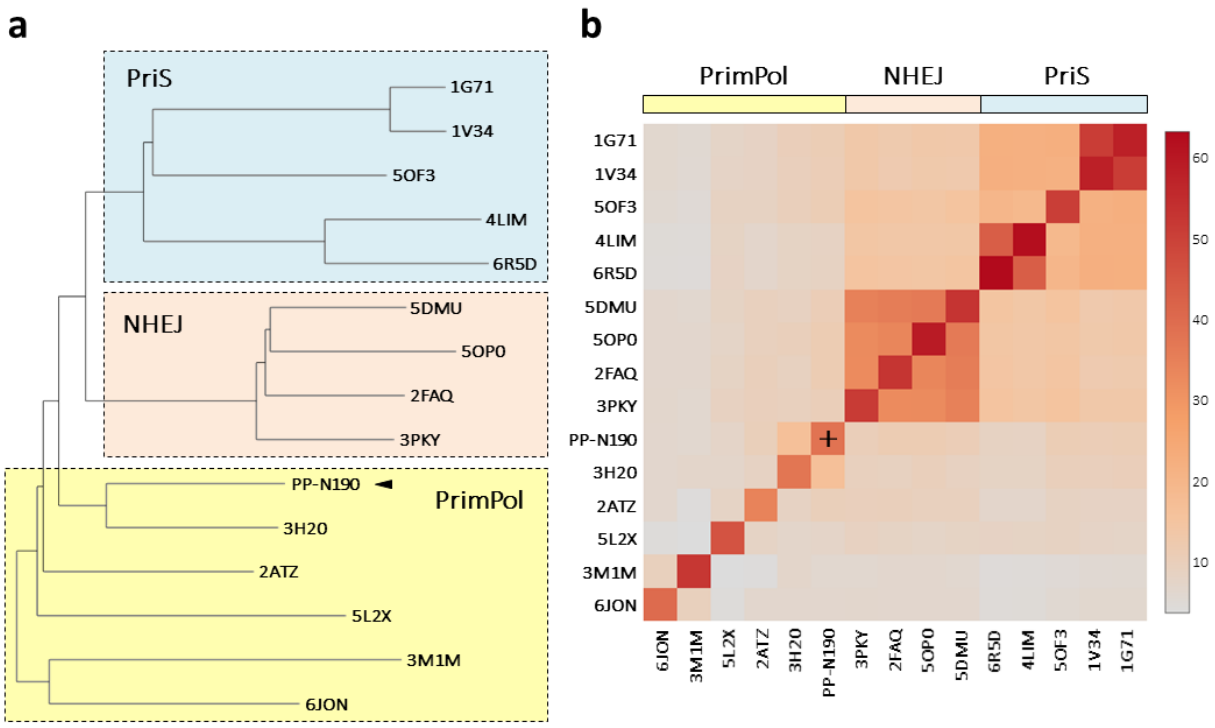
Supplementary Figure S4. Further tests on S-2L DatZ catalytic activity and its I22A mutant. The panels are constructed as in Fig. 4. **a.** HPLC analysis of nucleotides obtained after incubation of DatZ with dADP and dAMP, showing no discernible dephosphorylation products. **b.** The same for dATP, dZTP and dGTP incubated with DatZ I22A. Compared to the wild-type enzyme, the mutant shows reduced dATPase activity and improved, although still marginal, dZTPase activity. dATP to dA tri-dephosphorylation occurs in a single step.



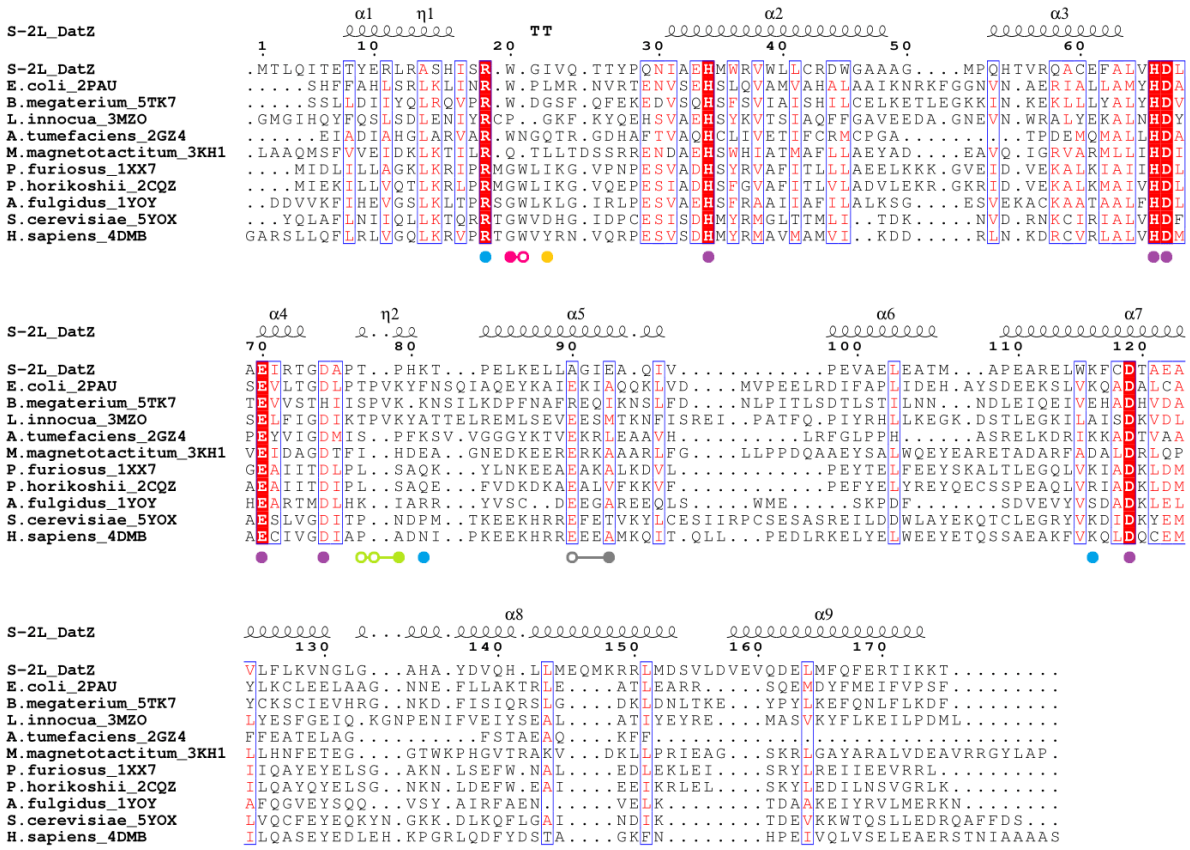
Supplementary Figure S5. Catalytic centre of S-2L DatZ dATPase with bound substrate, product and cofactors. Colour code as in Fig. 5b and Fig. 6a; residues K81 and K116, balancing the charge of the triphosphate, are also displayed. Water molecules and hydrogen atoms are omitted for clarity. **a.** Structure of DatZ with dA and Co^{2+}. The 2Fo-Fc electron density map around dA and Co^{2+} ions in the binding sites A and B is contoured at 1 sigma (black mesh). The anomalous signal at the wavelength of data collection is contoured at 10 sigmas (red mesh). **b.** Structure of DatZ with dATP and trace Zn^{2+}, using the same representation. The 2Fo-Fc electron density map around dATP and Zn^{2+} ion in the binding site A is contoured at 1 sigma (black mesh). Residual amounts of penta-coordinated Zn^{2+} can be identified by the anomalous signal at Zn edge, contoured here at 5 sigmas (red mesh).



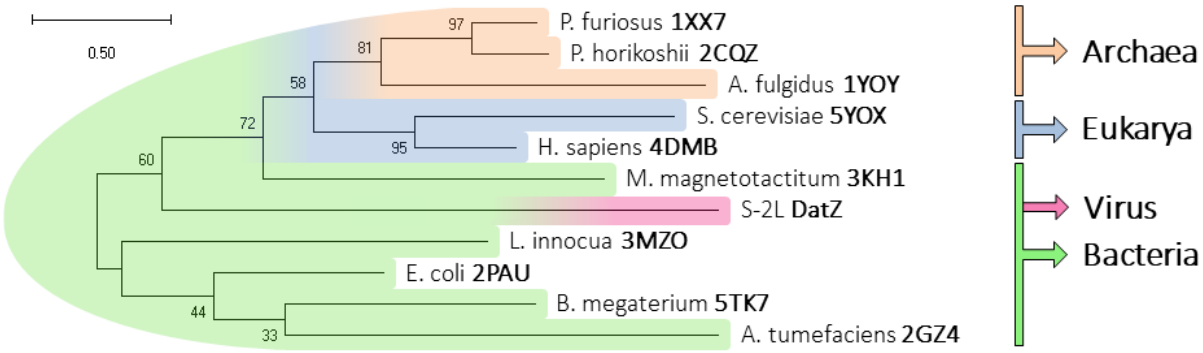
Supplementary Figure S6. **Sequence multialignment of close DatZ homologues co-occurring with *purZ* gene in related phages.** Numbering above the alignment refers to S-2L DatZ. Phages other than S-2L and ϕVC8 are described by their reference number (left). Dots below the alignment mark positions of residues crucial for DatZ: residues coordinating metal ions A and B (purple); residues stabilising the triphosphate (blue); W20 discriminating ribonucleotides (magenta); I22 providing steric hindrance for Z and G nucleobases (orange); P79 stabilising purine ring (lime). The remaining strictly conserved residues with hypothetical structural importance are marked by a brown dot.



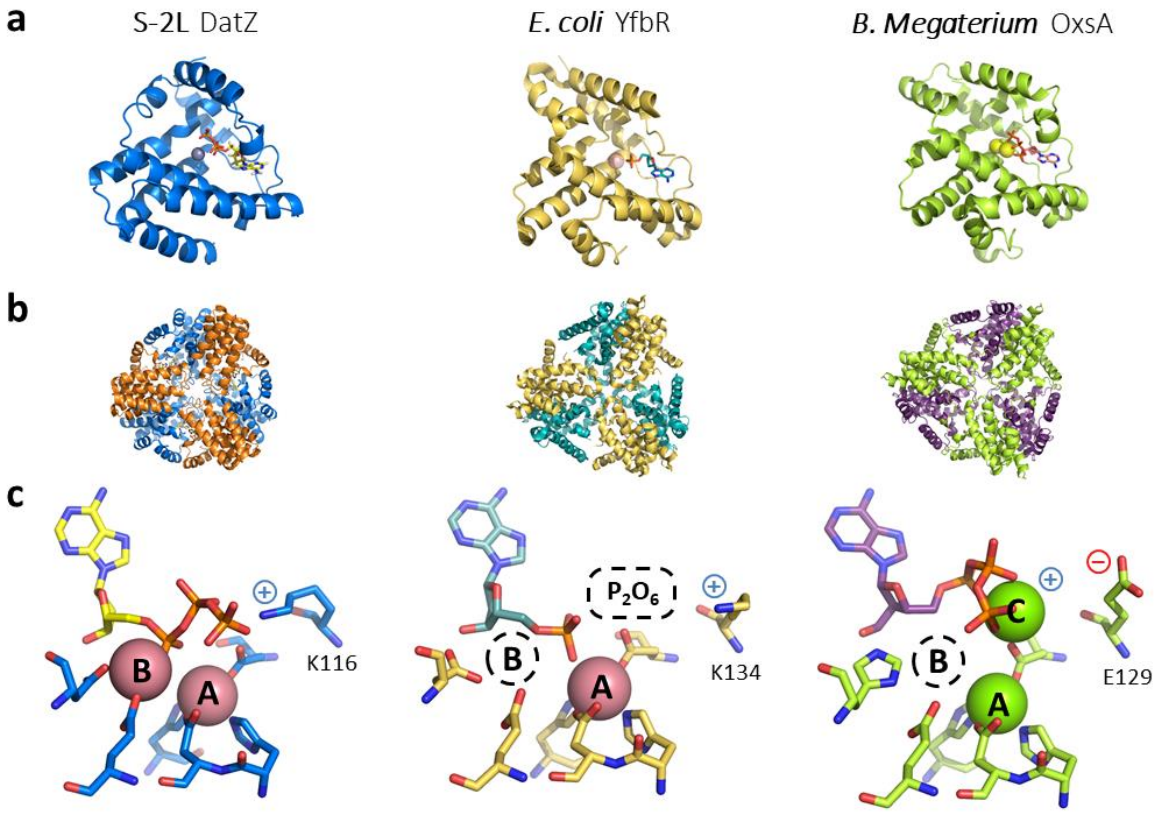
Supplementary Figure S7. **Structural classification of available AEP enzymes with Dali. a.** Dendrogram of AEP superfamily, derived by hierarchical clustering of the similarity matrix data. PDB codes are atop of the branches; PP-N190 is marked by a black triangle. Archaeo-eukaryotic PriS (light blue) and bacterial NHEJ primases (light orange) are monophyletic, and group together in so-called *AEP proper clade*. Primase-Polymerases (PrimPols, light yellow) are more diverge and spread-out across all three domains of life, viruses and plasmids. The AEP domain of S-2L PrimPol shares a recent ancestor with plasmidic RepB'. **b.** The similarity matrix data, with PDB codes to the left and below, and similarity scale to the right. PP-N190 is highlighted with a black cross. Each square represents with colours how close structurally a pair of AEP proteins is, from grey (no similarity) to dark red (high closeness).



Supplementary Figure S8. Structural multialignment of S-2L DatZ and all other HD phosphohydrolases whose crystal structure is available in the PDB. Organism names and PDB codes are indicated on the left. The significance of coloured dots below the alignment is as in Suppl. Fig. S6; the additional conserved residue E93 is marked by a grey dot. Empty circles highlight highly conserved residues with slightly shifted backbone positions with respect to S-2L (connected full circles), but with superimposed similar functional groups. Occasional unstructured and unbuilt regions in the middle were supplemented using sequence information alone while non-superposable N- and C-termini were ignored; in particular, an extended, but structured N-terminus of *L. innocua* and the last α-helix of *A. tumefaciens* that undergoes a considerable positional shift, were omitted.



Supplementary Figure S9. **Non-rooted maximum-likelihood phylogenetic tree of HD phosphohydrolases for all available molecular structures.** The tree was calculated using the alignment from the Suppl. Fig. S8. Enzymes divide into three groups: archaeal, eukaryotic and bacterial/S-2L, suggesting an acquisition of the *datZ* gene by S-2L's ancestor from a bacterium. The reference distance corresponds to an average 0.5 substitution per site. The topology of the bootstrap consensus tree is identical, supporting the result.



Supplementary Figure S10. **Common structural features across the HD phosphohydrolase family.** Comparison of S-2L DatZ with *E. coli* YfbR and *B. Megaterium* OxsA. **a.** Structures 6ZPC, 2PAU (chain A) and 5TK7. The overall protein fold is highly conserved, with RMSD of 2.0 Å and 2.3 Å, respectively, as well as the position of substrates and fixed catalytic divalent ions. **b.** Hexameric quaternary organisation of DatZ, YfbR and OxsA, extrapolated by using crystallographic symmetry operators to generate the hexamer. **c.** Comparison of the reaction centre between S-2L DatZ and *E. coli* or *B. megaterium* HD phosphohydrolases. DatZ is represented as in Fig. 6a; Yfbr E72A mutant structure with bound dAMP is taken from PDB 2PAU, with A72 residue swapped with natural E72 from PDB 2PAQ; OxsA is from PDB 5TK7. Features missing in YfbR and OxsA models can be inferred from DatZ structures (shown by dashed contours), suggesting that the metal ion site B is universal and has similar coordination across the whole protein family. Site C, observed for OxsA, would be a result of a switch from mostly conserved positively charged residue corresponding to DatZ's K116 to a negatively charged one, justifying the need for a third divalent cation.

Supplementary Table

Protein structure	**PrimPol-N190**	**DatZ + dA, Zn^{2+}**	**DatZ + dA, Co^{2+}**	**DatZ + dATP, Zn^{2+}**
PDB ID	6ZP9	6ZPA	6ZPB	6ZPC
Cell parameters				
Space group	P 1 21 1	R 3 2	R 3 2	R 3 2
a, *b*, *c* (Å)	59.2, 47.7, 66.2	141.5, 141.5, 53.6	141.9, 141.9, 53.7	141.7, 141.7, 53.7
α, *β*, *γ* (°)	90.0, 97.1, 90.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0
Data statistics				
Resolution (Å)	47.69 - 1.50 (1.54 - 1.50)	49.11 - 0.86 (0.87 - 0.86)	49.23 - 2.05 (2.10 - 2.05)	40.89 - 1.27 (1.30 - 1.27)
Wavelength (Å)	0.980097	0.729309	1.033202	0.980100
Rmerge (%)	10.2 (140.0)	10.5 (474.9)	9.2 (19.7)	6.6 (128.7)
Completeness (%)	96.8 (94.6)	100.0 (100.0)	99.6 (94.6)	99.8 (97.7)
Multiplicity	7.1 (7.0)	61.2 (51.8)	20.0 (17.4)	19.5 (17.6)
$I/\sigma(I)$	11.8 (1.4)	27.6 (1.7)	32.5 (17.2)	21.5 (2.0)
$CC_{1/2}$	0.999 (0.555)	1.000 (0.689)	0.998 (0.990)	1.000 (0.686)
Refinement				
Resolution (Å)	41.30 - 1.50	40.84 - 0.86	40.45 - 2.05	40.42 – 1.27
Unique reflections	56,610	172,368	13,094	54,250
R_{work}/R_{free} (%)	16.12/17.10	12.43/13.07	13.71/17.49	12.30/14.48
No. of non-hydrogen atoms				
Protein	2953	1449	1425	1435
Ligand	0	18	18	30
Metal ions	3	2	2	3
Water	552	219	237	231
Hydrogens added	No	Yes	No	Yes
Protein geometry				
Bond lengths (Å)	0.007	0.007	0.009	0.008
Bond angles (°)	0.85	1.23	0.91	1.31
Ramachandran favored/outliers (%)	100.00/0.00	99.42/0.00	99.42/0.00	99.42/0.00
Rotamers favored/poor (%)	95.81/0.32	96.82/0.00	98.04/0.00	98.06/0.00

B-factors (Å^2)				
Type	Anisotropic	Anisotropic	Isotropic	Anisotropic
Protein	19.44	11.64	15.20	17.87
Ligand	-	9.80	11.60	22.79
Metal ions	17.38	9.42	16.30	23.85
Water	34.99	25.26	25.17	33.52

Supplementary Table 1. Diffraction data collection and Model Refinement statistics. Numbers in parenthesis refer to the highest-resolution shell.

II. Complete metabolic pathway of 2-aminoadenine biosynthesis in a clade of *Siphoviridae* phages

Czernecki D, Kaminski PA and Delarue M

final stage of redaction

1. Preface

The next step expands upon the previous work done by P.-A. Kaminski and his colleagues in *Institut Pasteur* (Sleiman *et al*., submitted), completing the structure-function description of the main actors behind Z synthesis in cyanophage S-2L and identifying a set of 3 genes that we call the Z-cluster for ZTGC-DNA in phages.

All genomes of phages having a Z base replacing A in their DNA contain a homologue of adenylosuccinate synthetase (*purA*) that normally synthesizes the precursor of AMP from free L-aspartate, IMP and GTP as a source of energy (Eq. 1). In phages ϕVC8 and S-2L the product of this gene, called *purZ*, was shown to synthesize the direct precursor of 2-aminoadenine nucleotide from L-aspartate, dGMP and ATP (Eq. 2) (Sleiman *et al*., submitted to *Science*).

$$\text{PurA:} \quad \text{IMP} + \text{Asp} + \text{GTP} \rightarrow \text{sIMP} + \text{ADP} \tag{1}$$

$$\text{PurZ:} \quad \text{dGMP} + \text{Asp} + \text{ATP} \rightarrow \text{sdGMP} + \text{GDP} \tag{2}$$

These finding were supported by the crystal structures of ϕVC8 PurZ with its substrates. The product of PurZ, the deoxyguanylosuccinate nucleotide (sdGMP), is then reduced by the host's enzyme (PurB) to dZMP. Here, we solve the crystal structure of S-2L PurZ at 1.7 Å which suggests that, in addition to DatZ, PurZ can also act as a dATPase. We subsequently confirm this activity by in vitro catalytic tests and propose that it contributes to further deplete the host pool of nucleotides in dATP. Furthermore, *datZ* and *purZ* genes consistently cluster together in the genomes of a set of closely related phages from *Siphoviridae* family. We noticed that between datZ and purZ genes there exists a short gene conserved in all studied phages, similar to *mazG* in bacteria, that we call *mazZ*. We determine that its product, MazZ, acts as a diphosphohydrolase on dGTP, thus creating the dGMP substrate of PurZ in the 2-aminoadenine synthesis pathway. The crystal structure of MazZ tetramer with its reaction product was solved at 1.43 Å and provides structural evidence for an underlying two-metal-ion mechanism and a rationale for its specificity. These results complete our understanding of

2-aminoadenine metabolism across the whole viral clade (Fig. 26). The identification of a minimal Z-cluster made of these three genes (*datZ*, *purZ*, *mazZ*) could open the way for the successful establishment of a synthetic 2-aminoadenine pathway in other biological entities.

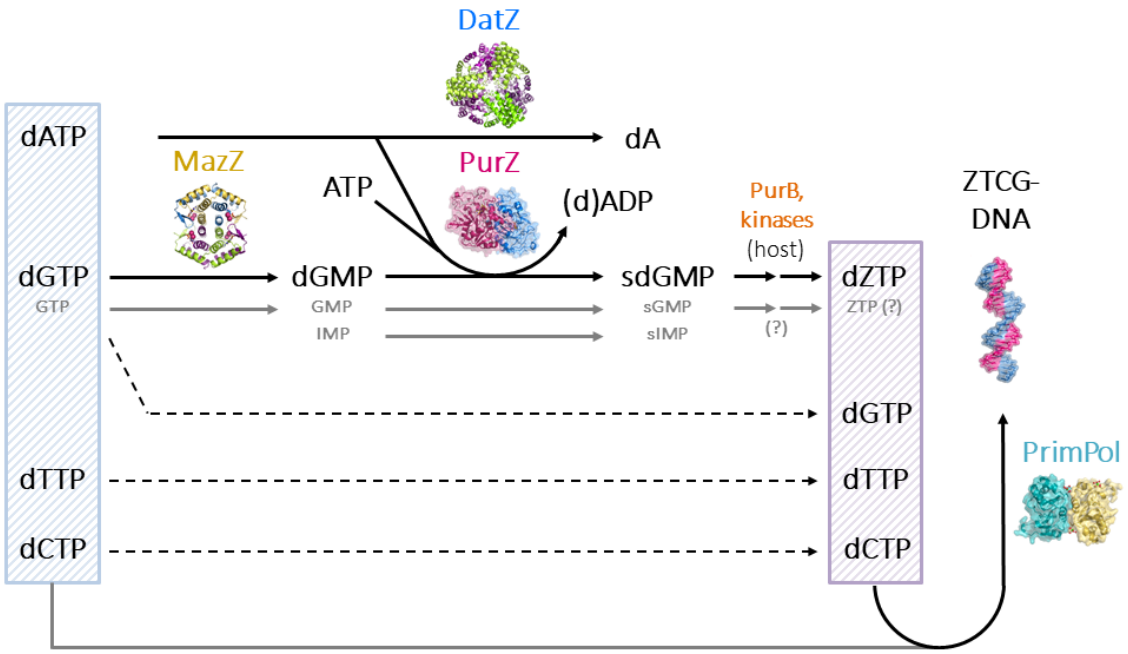


Figure 26. An overview of S-2L metabolic pathway. Side reactions performed by the enzymes are shown in gray; dashed arrows stand for no modification. The crystal structures of proteins solved in this thesis are shown in colors.

Similarly to PP-N190, the first crystals obtained for MazZ, which had the shape of triangular prisms (Fig. 27, left) were internally defective, despite their high reproducibility, fast growth, perfect shape in solution and good diffraction limit. Additionally, when exposed to cryoprotectant, either glycerol or ethanol, they start to degrade quickly. Luckily, new and faultless rod-shaped MazZ crystals grew in manually-reproduced drops of one of the conditions after several days – in these drops, the previous crystal form did not appear. As for the crystals of S-2L PurZ (Fig. 27, right), no major problems were encountered for structure solution and refinement, except that they would not directly co-crystallize with the nucleotide triphosphate and had to be soaked with it instead. Crystallographic conditions for all these crystalline forms are gathered in Table 5.

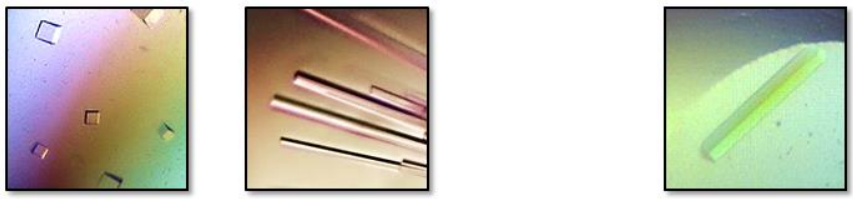


Figure 27. Defective prism-like and faultless rod-shaped crystals of MazZ (left); typical PurZ crystal (right).

Crystals	Conditions
PurZ (used)	• Tacsimate (15%), PEG 3350 (2% w/v), HEPES pH 7 (0.1 M)
MazZ (defective)	• Ca acetate (0.2 M), PEG 3000 (20% w/v), TRIS pH 7 (0.1 M) • Li_2SO_4 (0.2 M), $(NH_4)_2SO_4$ (1.26 M), TRIS pH 8.5 (0.1 M)
MazZ (used)	• Li_2SO_4 (0.2 M), $(NH_4)_2SO_4$ (1.26 M), TRIS pH 8.5 (0.1 M)

Table 5. Crystallization conditions for PurZ and MazZ crystals. One bullet point stands for one independent condition. All assays were done at 18°C, except the first one, done at 4°C.

Additionally, I collected diffraction data of PurZ bound to dGMP and ATP, in the same way that it was done with dATP. However, since it is redundant with data available for ϕVC8 PurZ (PDB ID: 6FM1) and of lower resolution (2.7 Å), the structure was not further refined past the preliminary stage (Fig. 28).

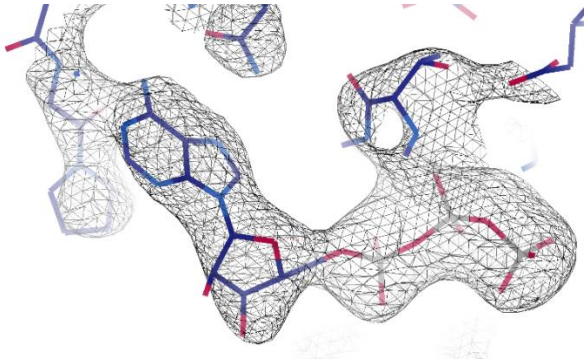


Figure 28. S-2L PurZ bound to dGMP and ATP (unrefined): the $2F_o$-F_c electron density around the ATP molecule is contoured at 1.8 sigmas.

Finally, I would like to discuss the possible regulation of the MazZ enzyme by the intermediates or product of the 2-aminoadenine pathway (sdGMP, dZMP or dZTP) as a hypothesis: indeed, when the nucleotide pool is complete in dZTP, there is no need to degrade any more dGTP that is otherwise necessary for DNA replication. For lack of time, these experiments are not included in the present manuscript, but they are planned to be done and all the necessary materials for that purpose are available.

Full metabolic pathway of 2-aminoadenine biosynthesis in a clade of *Siphoviridae* phages

Dariusz Czernecki (1,2), Pierre-Alexandre Kaminski (3) and Marc Delarue (1)*

(1) Unit of Structural Dynamics of Biological Macromolecules, UMR 3528 du CNRS, 25-28 rue du Docteur Roux, Institut Pasteur, 75015 Paris, France.

(2) Sorbonne Université, Collège Doctoral, ED 515, 75005 Paris, France

(3) Unit of Biology of Pathogenic Gram Positive Bacteria, Institut Pasteur, 75015 Paris, France.

Abstract

Several bacteriophages, such as the cyanophage S-2L, are known to have their genomic adenine (A) replaced completely by 2-aminoadenine (Z), thus providing a mechanism to avoid host DNA restriction endonucleases. The genomes of all these phages contain a homologue of adenylosuccinate synthetase (*purZ*), and in phages ϕVC8 and S-2L its product was indeed shown to synthesize a direct precursor of 2-aminoadenine from aspartate, dGMP and ATP. We previously showed that in phage S-2L a triphosphohydrolase (DatZ) is an important piece of the metabolism of Z incorporation in the phage DNA by cleaving dATP into dA, leaving dZTP as the sole thymine-matching substrate for the otherwise non-specific DNA polymerase. Here we solve the crystal structure of S-2L PurZ at 1.7 Å which suggests that, in addition to DatZ, PurZ can also act as a dATPase. We confirm this activity by catalytic tests, proving that it contributes to dATP pool depletion. Furthermore, we now show that between *datZ* and *purZ* genes, clustered together in the genomes of related *Siphoviridae* phages, there exists a third, conserved short gene. Using S-2L's revisited genome sequence, we express this gene and determine that its product, called MazZ, acts as a diphosphohydrolase on dGTP, thus creating the dGMP substrate of PurZ in the 2-aminoadenine synthesis pathway. The crystal structure of MazZ tetramer with its reaction product at 1.43 Å provides structural evidence for an underlying two-metal-ion mechanism and a rationale for its specificity. These results complement our understanding of 2-aminoadenine metabolism across the whole viral clade, and could open the way for the successful establishment of a synthetic 2-aminoadenine pathway in other biological entities.

INTRODUCTION

At least since the last universal common ancestor (LUCA), only four nucleobases – adenine (A), guanine (G), cytosine (C) and thymine (T) or its analogue, uracil (U) – are used to encode genetic information in DNA or RNA polymers, respectively, and their metabolic pathway is conserved across all branches of cellular life (1). This principle can be extended to the vast majority of smaller biological agents, such as viruses, which are important agents of evolution (2). Despite that, a genetic material may be subject to many natural nucleobase modifications. They either constitute an additional, epigenetic information, or exist as a consequence of an arms race with restriction-modification systems in the hosts. Some well-described viral examples of the last behavior are represented by double-stranded DNA (dsDNA) bacteriophages from the order *Caudovirales*: for instance, phages T2, T4 and T6 systematically substitute 5-hydroxymethylcytosine (5hmC) for cytosine (3), whereas phage 9g contains archaeosine (G+) which replaces a quarter of its genomic guanine (4), enabling its DNA to resist 71% of cellular restriction enzymes (5).

Although numerous nucleobase modifications exist, such alterations are made almost exclusively without changing the Watson-Crick base-pairing scheme. The only known exception to this rule was made apparent by the discovery of cyanophage S-2L, belonging, like the aforementioned phage 9g, to the family *Siphoviridae* (6). S-2L abandons the usage of genomic adenine in favor of 2-aminoadenine (2,6-diaminopurine or Z), which has a supplementary amino group in position 2 of the purine ring (Fig. 1A). The resulting ZTGC-DNA of the cyanophage has an improved thermal stability with respect to classical ATGC-DNA (7, 8) and proves to be almost completely resistant to adenine-targeting restriction enzymes (9).

The metabolic pathway of 2-aminoadenine synthesis in S-2L and related *Siphoviridae* phages was proposed in a recent study (Sleiman *et al*., submitted). This work identified the key enzyme of Z metabolism as PurZ, a homologue of adenylosuccinate synthetases (AdSS encoded by *purA* gene), which creates the immediate precursor of

2-aminoadenine monophosphate, deoxyguanylosuccinate monophosphate (sdGMP), from standard dGMP, free aspartic acid (Asp) and ATP as an energy donor. The activity of PurZ was tested *in vitro* for cyanophage S-2L and vibriophage ϕVC8, and the crystal structure of the latter was reported with these substrates (Sleiman *et al*., submitted). Finally, the enzymes necessary for the subsequent reactions generating dZTP, such as PurB reducing the N6-substituted position to an amine group, are not encoded on the phages' genomes: it was postulated that they were provided by their hosts. Indeed, the corresponding enzymes of *V. cholerae* showed a relaxed purine specificity and thus complemented the pathway of dZTP metabolism of ϕVC8 *in vitro*.

In parallel to that, our previous work (Czernecki *et al*., submitted) described a dATP-specific triphosphatase (DatZ) which eliminates dATP from the pool of available dNTPs substrates for the otherwise non-discriminative DNA primase-polymerase (PrimPol) of cyanophage S-2L. Through structural studies of both enzymes, we proposed a mechanistic rationale for their activities and specificities, or lack thereof. Based on the co-conservation of both *purZ* and *datZ* genes, we expanded the idea of dATP depletion to related ZTGC-DNA-containing viruses.

In catalytic assays, PurZ enzymes were tested with both classical or logical substrates: dGMP, GMP or IMP on one hand, and ATP or GTP on the other (Sleiman *et al*., submitted). Despite that, close inspection of available ϕVC8's structure (PDB ID: 6FM1) suggests a possible lack of selection on 2' hydroxyl group of the ATP's ribose, enabling the potential usage of dATP as an energy donor. The unusual catalytic activity of PurZ may ultimately lead to potential applications in the enzymatic synthesis of unnatural purines from standard precursors: for instance isoguanine, a component of 8-letter synthetic nucleic acids (10), may be generated by using natural xanthosine monophosphate (dXMP) as a substrate.

Here we broaden the substrate spectrum of S-2L's PurZ by solving its structure in a complex with dGMP as a substrate and dATP as an alternative energy donor. We confirm the alternative role of PurZ as a dATPase, that can nevertheless stay active after the ultimate pool depletion, using ATP. Moreover, through close inspection of all

closely related *Siphoviridae* phages, we identify that two (interchanging) gene variants of MazG-like nucleotide pyrophosphatase (MazZ) compose the third and final element of a conserved cluster, which we call the Z-cluster. We expose the specificity of S-2L's MazZ for both guanosine and deoxyguanosine triphosphates, leading to GMP or dGMP, which places the enzyme directly upstream of PurZ in the 2-aminoadenine synthesis pathway. By resolving the crystal structure of MazZ bound to the first dephosphorylation product of dGTP, we identify crucial residues important for the enzyme's activity. These discoveries tie up pre- and post-infection triphosphate pools and complete the functional and structural description of 2-aminoadenine metabolism in the cyanophage S-2L model.

RESULTS

Cyanophage S-2L genome and its relatives: a conserved Z-cluster

We resequenced the genome of phage S-2L, revealing that it is in fact circular (Fig. 1B). We manually verified all 187 possible ORFs identified by ORFfinder (11) for possible protein homologues in BLAST (11) and conserved motifs in MOTIF (12). We identified 41 genes distributed evenly across the DNA sequence, rarely overlapping (and if so, with the very ends), with a total length equivalent to 75.7% of the genome. They are partitioned into 3 functional blocks, or classes: early short genes, replication-related middle ones and late genes encoding for structural proteins, with lysozyme at the very end. This modularity is similar to other phages from *Siphoviridae* family (13, 14), the family to which cyanophage S-2L belongs, including the model phage λ (15). We found no Shine-Dalgarno (SD) motifs upstream the genes, which is consistent with low SD motif conservation in cyanobacteria (16, 17), closely mimicked by cyanophages (18). The elevated GC content (69.4%) is fairly constant across the whole genome. Interestingly, the S-2L genome can be divided into two parts of roughly equal length, where almost all genes follow only one direction of translation (Fig. 1B).

Genes from class II were of obvious interest for studying the peculiar metabolism behind S-2L DNA synthesis, as they translate to a number of important replication-related proteins. As identified in our previous study (Czernecki *et al*., submitted), they consist of: single-strand DNA-binding (SSB) protein, VRR-NUC flap endonuclease, exonuclease VIII, superfamily II helicase, deoxyguanidylosuccinate synthetase (PurZ), HD phosphohydrolase (DatZ) and DNA primase-polymerase from AEP family (PrimPol). Additionally, during S-2L genome resequencing we unravelled the presence of a third gene placed tightly between *purZ* and *datZ*, whose product corresponds to a MazG-like pyrophosphatase: for naming consistency, we will hereafter call it MazZ.

Next, we compared S-2L with its close relatives bearing the crucial 2-aminoadenine-linked enzymes, available through NCBI BLAST (11) and TARA BLAST (19) interfaces. Aside from S-2L, we identified 15 phage sequences of full or major genome coverage. On Fig. 1C we show the replication-related genome sections of these phages, along with their phylogeny constructed on PurZ and DatZ proteins. We could distribute all phages into four clusters of similar replication-related genome architecture: until now, cyanophage S-2L is the only representative of its clade, clearly distinct from others. We observe that *mazZ* gene is always co-conserved with *datZ* and *purZ*, although in one of two possible isoforms. MazZ-1, the isoform of ϕVC8, is replaced with MazZ-2 in clade S-2L and the immediately related clade A, as well as in one singular case of a phage from clade B. MazZ-2 in clade A also seems to undergo a fusion with a HNH domain and, in another instances, with a different, unknown domain.

No other gene seems to be closely conserved with the three aforementioned ones. Although there is a number of accompanying genes common for clades A, B and C, they are absent in cyanophage S-2L. We refer therefore to the closely connected *datZ*, *mazZ* and *purZ* genes as the 2-aminoadenine cluster, or Z-cluster. Further, based on recent studies on ϕVC8 and S-2L and the present work, we argue that this cluster is the smallest element necessary and sufficient for complete ATGC- to ZTGC-DNA conversion in a phage.

S-2L PurZ alternate dATPase activity and secondary monophosphate substrates

In order to expand on previous results concerning PurZ, we have cloned and overexpressed a synthetic version of S-2L's gene in *E. coli*. After purification of the product, we performed catalytic tests with various nucleotide substrates, and followed the appearance of the products by HPLC (Fig. 2, Suppl. Fig. S1). Structural analysis of ϕVC8's close homologue (PDB ID: 6FM1) suggested to compare the reaction time-course with either dATP or ATP as the phosphate donor (Fig. 2A). The result shows no specificity of PurZ towards the *ribo*- and *deoxy*- variants: the molecular nature of this lack of discrimination and its importance for the phage are discussed in the following sections. Furthermore, we found that with the appropriate energy donor PurZ is also able to catalyze the reaction with GMP and IMP as substrates, though with decreasing efficiency compared to dGMP (Suppl. Fig. S1).

As a control, we confirmed the previously reported activity of PurZ, as well as the *E. coli* adenylosuccinate synthetase (AdSS) (20) (Sleiman *et al*., submitted) in the same experimental setup (Fig. 2B).

Structure of S-2L PurZ with bound dGMP and dATP at 1.7 Å resolution

We crystallized S-2L PurZ with two of its substrates, dGMP and dATP, and solved its structure at 1.7 Å resolution, using the related structure of ϕVC8 (PDB ID: 6FM1) as a template in molecular replacement (Table 1, Fig. 3A). In accordance with the 41% sequence identity between the two proteins, their structures are highly similar, with an RMSD of 1.6 Å. Both ligands have exactly the same binding mode as previously described (Fig. 3B): notably, dATP closely superposes with ATP, with virtually no difference in the position of the 2' carbon atom.

Enzymes from AdSS family are known to be functional dimers (21, 22) or even tetramers (23). Even though there is one molecule of PurZ in the crystallographic asymmetric unit, a dimer can be reconstructed by using a crystallographic two-fold axis (Fig. 3C). As expected, there is a large surface interaction area of 2488 $Å^2$ between the pair of proteins. Although the enzyme crystallized in a different space group from

the one observed in ϕVC8 enzyme's crystals, this observation can be extended to the original PurZ structure as well: analysis of crystal contacts between symmetry-related molecules with PISA indicates indeed the presence of a stable dimer.

Very low B-factors (Fig. 3D), especially on the interface, suggest an exceptional overall rigidity of the dimer; the only exception is the Y272-L278 loop above the reactants. This flexible loop corresponds to the loop 299-304 in *E. coli* AdSS, shown to be ordered only in the presence of a non-hydrolyzable analogue of aspartate, hadacidin (thus referred to as "the aspartate loop") (24). The tip of the aspartate loop (T273-T276) is completely conserved between S-2L and ϕVC8, both in sequence and structure. The signature G273T mutation for the PurZ clade is possibly important for the activity (Sleiman *et al*., submitted).

<u>dGMP-binding site</u>

Residues G22, S23, N49, A50, T132, Q190, V241 and the backbone atoms of Y21 form a pocket where the base moiety of dGMP is placed. Y21-S23 are positioned next to the amino group of dGMP in position 2, which is absent in inosine. The sidechain of S23 in the vicinity of this amino group, specific to PurZ sequences, is likely to be responsible for the guanine specificity (Sleiman *et al*., submitted). Its negative partial charge is further stabilized by the close R280 sidechain. Residues G130, S131, R146, Y203, C204, T205 and R240 complete the dGMP pocket. R146 is protruding from the dimer's other molecule, forming an ion pair with the negative charge of the α-phosphate, a feature already described for *E. coli* AdSS as well (25). V241 is ideally placed to sterically interfere with the ribose 2'-OH group in ribonucleotides, establishing preference towards the substrate in the *deoxy*- form. Although it is conserved in both AdSS and PurZ, its position is shifted in purZ: the reason behind this discrepancy is examined in the discussion section.

(d)ATP -binding site

Just like ATP in the PurZ structure of ϕVC8, the base moiety of dATP is stabilized by stacking interactions with F306 and P336 residues. Oxygen atoms from the side-chains of N305, Q308 and the backbone of G335 form hydrogen bonds with adenine's amino group in position 6. Importantly, the amino group of Q308 would destabilize the amino group on C2 of bases G and Z; the same effect seems to take place with the amino group of N297 in ϕVC8. Identically to its homologue, the rest of the ligand interacts with S-2L PurZ almost exclusively through its triphosphate tail with residues S23, G25, G51, H52 and T53; only the last residue is also in contact with the deoxyribose moiety, touching it from the C3' atom side. The position inferred for the 2'-OH of ATP would be entirely exposed to the solvent, identically to what is seen for ϕVC8 (Sleiman *et al*., submitted) (PDB ID: 6FM1). This finding confirms the lack of a selection mechanism for ribose/deoxyribose variants across PurZ enzymes.

S-2L MazZ: function and structure of (d)GTP pyrophosphatase

To complete our studies on the Z-cluster in cyanophage S-2L, we sought to investigate the catalytic properties of MazZ. We cloned and overexpressed its gene, purified the enzyme and subjected it to HPLC tests.

The results are shown on Fig. 4. As expected from a member of the all-α NTP pyrophosphohydrolase (NTP-PPase) superfamily, MazZ is able to remove two terminal phosphates from a nucleotide triphosphate. Its preferential substrates were identified to be dGTP and GTP, which the protein rapidly dephosphorylates to dGMP and GMP, respectively. In contrast, it shows no significant activity with dATP and dZTP: trace of activity in dATP dephosphorylation starts to be visible only with incubation times 8 times longer. Thus, S-2L MazZ evidently exerts strong discrimination on base (purine) moiety, but seemingly none on the *ribo*- and *deoxy*- nucleotide variants (sugar moiety).

We successfully obtained crystals of MazZ and solved its structure, bound to the dephosphorylation product of dGTP and catalytic Mn^{2+} ions, at 1.43 Å resolution (Table 1, Fig. 5), by making use of Mn^{2+} anomalous signal for *ab initio* structure

determination. Contrary to other members of the all-α NTP-PPase superfamily, that typically contain only four to five α-helices, each MazZ protein chain has two additional β-strands on its C terminus (Suppl. Fig. S4). Together, they form a homotetramer – a dimer of two tight dimers – with four identical active sites, each formed by two intertwining chains of the tight dimer. The whole tetramer constitutes the asymmetric unit, with very little deviations between the monomers (RMSD of 0.2-0.6 Å). The only noticeable differences in electron density between the chains lie in the solvent-exposed D43-H46 flexible loop and on both N- and C-termini, partly influenced by crystallographic contacts.

The electron density of four dGDP nucleotides, occupying each of the catalytic pockets, revealed the presence of two phosphate groups (Fig. 5B). It signifies that under the crystallization conditions the enzyme crystallized after having reduced dGTP to dGDP, but before reducing the latter to dGMP. The nucleotides are placed in very tight pockets, engulfed by the enzyme from every side except from the solvent-exposed β-phosphate. There, three catalytic Mn^{2+} ions are found, as indicated by a strong anomalous signal.

<u>dGDP -binding site</u>

Guanine nucleotides are completely sandwiched by the protein. From one side, the ligand is held by residues I12, W15, I16, N20, K31, E35, D53, I56, L57, and D60 of one chain. The second chain of the tight dimer completes the pocket from the other side with residues K76, N80, W85, A92, M93, R94 and H95.

Looking at the essential guanine functional groups, the closest residue to the oxygen atom on C6 is N20 through its amide nitrogen at 3.2 Å; on the other hand, the amino group on C2 is only 3.0 Å away from the carboxyl group of D60. Both of these hydrogen bond interactions are completed with hydrophobic interactions with the purine ring. Hence, the specificity of MazZ towards guanine emerges from the cavity volume matching its shape and charge compatibility with the two essential chemical groups of guanine.

We observe no steric hindrance for the possible presence of the ribose 2'-OH group. Mutation of the closest I56 could potentially decrease the specificity for ribonucleotides; however, it also contacts the base's 2-amino group and is surrounded by an intricate network of other residues. Thus, improving the specificity of MazZ towards deoxyribose is probably not a trivial task.

Catalytic Mn^{2+} ions

The three Mn^{2+} ions place themselves at the pocket's opening: one between the α- and β-phosphates of dGDP, one on the opposite side of the β-phosphate, and one next to where the γ-phosphate would extend. We name them ions A, B and C, respectively; they are strictly equivalent to the Mg^{2+} ions found in *C. jejuni* dimeric dUTPase (PDB 1W2Y), another member of the superfamily (26). The ion A is coordinated by residues E34, E35 and E38; ion B by the E50 and D53; and ion C by E38 and E50. With oxygen atoms from phosphates and water molecules filling the coordination shells, these ions are all hexa-coordinated, as seen in the *C. jejuni* dUTPase. The two-metal-ion has already been described for two other distinct representatives of the all-α NTP pyrophosphohydrolase superfamily (26, 27); thus, we expect this mechanism to be valid for MazZ's activity as well. However, we propose that the dephosphorylation occurs with two independent cleavage steps, which is consistent with the presence of three catalytic ions and an intermediate diphosphate in MazZ structure. Additionally, residue R83, positioned only 2.8 Å away from the β-phosphate, is probably important for the excision step by stabilizing the penta-coordinated intermediate that appears during two-metal-ion catalysis (28).

The complete 2-aminoadenine biosynthetic pathway

Combining the information above with previous investigation of S-2L and ϕVC8 homologous enzymes (Czernecki *et al*., submitted; Sleiman *et al*., submitted), we now propose a complete 2-aminoadenine biosynthesis pathway for cyanophage S-2L (Fig. 6). Upon infection by the phage, the available cellular pool of dNTPs is heavily modified

by the three conserved enzymes composing the Z-cluster. On one hand, dATP is completely eliminated by dephosphorylation to dA by DatZ or to dADP by PurZ. On the other hand, MazZ uses a fraction of the dGTP pool to make dGMP, further transformed by PurZ to sdGMP, the direct precursor of deoxy-2-aminoadenine monophosphate (dZMP).

The finishing steps of dZTP synthesis are carried out by non-specific host enzymes participating in dATP production, as their genes are absent in the S-2L's genome. It is conceivable that the production of these non-specific enzymes is additionally upregulated by the infected host cell sensing and fighting the depletion of its dATP reserve. The final composition of the dNTP pool thus consists of dATP being replaced by dZTP, and the Primase-Polymerase of S-2L, non-specific to A and Z, readily inserts the surrogate base in front of any instructing thymine.

We conclude that other phages with conserved Z-cluster undergo similar mechanism of dATP removal and dZTP production and incorporation, whether their respective DNA polymerases are more prone to the formation of the Z:T pair or not (Pezo *et al*., in preparation).

Discussion

Molecular reasons for substrate selectivity in PurZ vs AdSS

With Dali (29), we found every AdSS family member available in the PDB. Using PROMALS3D (30), we extracted the existing structural information in order to construct a reliable structurally-informed sequence multialignment for both AdSS and viral PurZ (Suppl. Fig. S2). 36 residues are strictly conserved, while 26 further residues have very little variability (2 similar variants or only occasional mutations), making up for 17.3% of S-2L PurZ total length. These two classes cluster in the catalytic pocket and the surrounding layer, respectively (Suppl. Fig. S3A). We note that several conserved residues seemingly important for the ternary structure (P256, F283, D18) maintain the same physical position of their sidechain, but are subjected to some

sequence rearrangement in S-2L and ϕVC8. Finally, 16 residues are unique to the viral branch, otherwise strictly conserved in cellular AdSS. Their placement is intermediate compared to previous classes, although several such residues (S23, T273 and Q308) interact with the substrates.

The position of the guanidino group of S-2L PurZ R244 corresponds to ϕVC8 PurZ K267, whose location on the backbone is shared with the otherwise conserved Arg residue (R303 in *E. coli*). However, whereas in AdSS this arginine balances the partial charge of O2' of the ribose ring of IMP at 2.5-4 Å distance (PDB IDs: 1P9B, 5I34, 5K7X) and interacts with the free aspartate substrate (31), in viruses these residues extend noticeably further, being 7.9 Å away from C2' in S-2L and 8.5 Å in ϕVC8 (PDB ID: 6FKO), precluding any possible stabilization of the ribonucleotide. Moreover, residue V241, mentioned above as important for ribose discrimination, is also present in standard AdSS family representatives. However, in phages' PurZ an insertion deforms the loop that contains it, pulling it slightly closer to the C2' ribose atom – from 4.3-5.3 Å as seen in AdSS (PDB IDs: 1P9B, 2DGN, 2GCQ, 4M9D, 5I34, 5K7X) to 3.7 Å (S-2L) and 3.6 Å (ϕVC8).

Lastly, there are two large deletions specific to the phages and *P. horikoshii* involving a helix-loop-helix motif and a strand-strand-helix-strand structural element (Sleiman *et al*., submitted). They are both solvent-exposed and do not appear to intervene in the catalytic activity or protein dimerization. In contrast, S-2L PurZ has a unique additional C-terminal helix α11, that partially compensates the second deletion. We mapped these indels onto S-2L and *E. coli* structures (Suppl. Fig. S3B). Using the structure-informed sequence alignment, we constructed a phylogenetic tree (Suppl. Fig. S3C) that supports the hypothesis on PurZ's archaeal origin (Sleiman et al.).

Classification and function of MazG-HisE proteins

Closely related MazG and HisE proteins share an evolutionary history with dimeric dUTPases, and all three constitute the all-α NTP pyrophosphohydrolase (NTP-PPase)

superfamily (32). The basic unit of these enzymes is a four or five α-helical chain, with only one notable exception to this date: *A. fulgidus* AF_0060 (MazG-like) protein, where an extended α_1 helix structurally compensates for the lack of α_4 helix, despite its opposite directionality (PDB 2P06). In MazG and HisE enzymes, believed to represent the ancestral fold, two of these 4-helical chains intertwine, forming a tight dimer with two symmetric catalytic sites made from both subunits. Sometime later in evolution, the ancestor of dimeric dUTPases underwent gene duplication and fusion, creating a covalent equivalent of such a dimer that subsequently lost one redundant catalytic site (32). Most MazG and HisE proteins, like S-2L MazZ, dimerize further through a hydrophobic surface, forming a tetramer with four active sites; dimeric dUTPases accordingly dimerize as well, in a geometrically similar arrangement (32). For a boundary case, HisE of *M. tuberculosis* assembles into a loose tetramer with comparatively small surface area between the two dimers, present however in two independent crystalline forms (33). Lastly, *E. coli*'s MazG do not form a tetramer at all – this however may be a consequence of its special form, where two independent domains with different activities are fused, dimerizing into two tight dimers joint by flexible linkers (34).

To our knowledge, in spite of a number of determined MazG and HisE structures, it is the first time that their ternary similarity and quaternary structure identity is recognized (Suppl. Fig. S4). Thus, in absence of distinct structural and functional features justifying the separation of these families, we will refer to all representatives of this fold as "MazG-HisE" enzymes. With the strong structural support described in this work, we may now include S-2L MazZ in this group.

HisE members are closely related and are involved in bacterial histidine synthesis pathway as a phosphoribosyl-ATP pyrophosphatase (35), often fusing with HisI that catalyzes the following reaction in the pathway (36). On the other hand, MazG proteins play house-cleaning or related functions, like degrading an alarmone (p)ppGpp (37) or aberrant dUTP (38). Varying specificity of all MazG-HisE enzymes is reflected in the divergence of residues in contact with the ligand – only the catalytic residues and fold-

related hydrophobic ones are consistently conserved across the superfamily (32). As noted before, these active site residues were identified to coordinate Mg^{2+} ions in the two-metal-ion mechanism similar between MazG-HisE and dimeric dUTPases (26, 27).

Comparison of S-2L MazZ with other MazG-HisE representatives

Using Dali (29) we identified all PDB structures of MazG-HisE enzymes, 15 being available at the time of writing this article. However, as the protein chains are short and divergent, even the structure-guided sequence multialignment is not sensitive enough to unravel relations between the representatives with a satisfactory level of confidence. Similarly, Dali's structural clustering did not find strong structural connections concerning the protein of interest, apart from regrouping all HisE enzymes together – S-2L MazZ clearly stays outside of this group. Indeed, together with previously undescribed secondary structure extensions, these findings confirm the structural uniqueness of S-2L MazZ.

As expected from the high divergence, very few residues are conserved between S-2L MazZ and these representative sequences. Apart from catalytic negatively-charged residues coordinating Mg^{2+} ions conserved almost unanimously, some loose conservation only concerns two electropositive residues (K31 and R83) interacting with the phosphates. This lack of information does not allow to derive any phylogenetic relations between S-2L MazZ and other MazG-HisE proteins. Between the few homologues of the same MazZ-2 variant in other phages bearing the 2-aminoadenine cluster, almost all of the crucial residues are well conserved; the only dissenting ones are I16 and N20, exclusive to S-2L.

BLAST searches indicate that the closest non-viral homologues of S-2L MazZ are in great majority bacterial, mainly from the phyla Terrabacteria and Proteobacteria.

Finally, the fact that both MazZ and PurZ of cyanophage S-2L show side reactions with ribonucleotide variants of their preferential substrates may signify parallel synthesis of sGMP. As PurB does not seem to differentiate between ribo- and

deoxyribonucleotides (39), it is possible that during the phage infection GTP is transformed to ZMP, which may then be phosphorylated to ZTP and incorporated into the nascent mRNA molecules. However, with a large ATP pool and in the absence of efficient ATPase, the A-to-Z substitution ratio in RNA would be expected to be drastically smaller, with little to no physiological effect on both transcription and translation.

Bibliography

1. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology* **1**, 127–136 (2003).

2. P. Forterre, The origin of viruses and their possible roles in major evolutionary transitions. *Virus Research* **117**, 5–16 (2006).

3. G. R. Wyatt, S. S. Cohen, The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine. *Biochemical Journal* **55**, 774–782 (1953).

4. J. J. Thiaville, *et al.*, Novel genomic island modifies DNA with 7-deazaguanine derivatives. *PNAS* **113**, E1452–E1459 (2016).

5. R. Tsai, I. R. Corrêa, M. Y. Xu, S. Xu, Restriction and modification of deoxyarchaeosine (dG +)-containing phage 9 g DNA. *Scientific Reports* **7**, 8348 (2017).

6. M. D. Kirnos, I. Y. Khudyakov, N. I. Alexandrushkina, B. F. Vanyushin, 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **270**, 369 (1977).

7. I. Ya. Khudyakov, M. D. Kirnos, N. I. Alexandrushkina, B. F. Vanyushin, Cyanophage S-2L contains DNA with 2,6-diaminopurine substituted for adenine. *Virology* **88**, 8–18 (1978).

8. M. Cristofalo, *et al.*, Nanomechanics of Diaminopurine-Substituted DNA. *Biophysical Journal* **116**, 760–771 (2019).

9. M. Szekeres, A. V. Matveyev, Cleavage and sequence recognition of 2,6-diaminopurine-containing DNA by site-specific endonucleases. *FEBS Letters* **222**, 89–94 (1987).

10. S. Hoshika, *et al.*, Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* **363**, 884–887 (2019).

11. D. L. Wheeler, *et al.*, Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**, 28–33 (2003).

12. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya, The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42–46 (2002).

13. H. Brüssow, F. Desiere, Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Molecular Microbiology* **39**, 213–223 (2001).

14. J. Murphy, *et al.*, Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci Rep* **6**, 1–13 (2016).

15. H. Echols, H. Murialdo, Genetic map of bacteriophage lambda. *Microbiology and Molecular Biology Reviews* **42**, 577–591 (1978).

16. J. Ma, A. Campbell, S. Karlin, Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures. *Journal of Bacteriology* **184**, 5733–5745 (2002).

17. S. Nakagawa, Y. Niimura, K. Miura, T. Gojobori, Dynamic evolution of translation initiation mechanisms in prokaryotes. *PNAS* **107**, 6382–6387 (2010).

18. Y. Wei, X. Xia, Unique Shine–Dalgarno Sequences in Cyanobacteria and Chloroplasts Reveal Evolutionary Differences in Their Translation Initiation. *Genome Biol Evol* **11**, 3194–3206 (2019).

19. A. Priyam, *et al.*, Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol Biol Evol* **36**, 2922–2924 (2019).

20. I. Lieberman, W. the technical assistance of W. H. Eto, Enzymatic Synthesis of Adenosine-5'-Phosphate from Inosine-5'-Phosphate. *J. Biol. Chem.* **223**, 327–339 (1956).

21. W. Wang, A. Gorrell, R. B. Honzatko, H. J. Fromm, A Study of Escherichia coli Adenylosuccinate Synthetase Association States and the Interface Residues of the Homodimer. *J. Biol. Chem.* **272**, 7078–7084 (1997).

22. R. Jayalakshmi, K. Sumathy, H. Balaram, Purification and Characterization of Recombinant Plasmodium falciparum Adenylosuccinate Synthetase Expressed in Escherichia coli. *Protein Expression and Purification* **25**, 65–72 (2002).

23. S. Mehrotra, H. Balaram, Kinetic Characterization of Adenylosuccinate Synthetase from the Thermophilic Archaea Methanocaldococcus jannaschii. *Biochemistry* **46**, 12821–12832 (2007).

24. R. B. Honzatko, H. J. Fromm, Structure–Function Studies of Adenylosuccinate Synthetase from Escherichia coli. *Archives of Biochemistry and Biophysics* **370**, 1–8 (1999).

25. B. W. Poland, *et al.*, Crystal structure of adenylosuccinate synthetase from Escherichia coli. Evidence for convergent evolution of GTP-binding domains. *J. Biol. Chem.* **268**, 25334–25342 (1993).

26. O. V. Moroz, *et al.*, The Crystal Structure of a Complex of Campylobacter jejuni dUTPase with Substrate Analogue Sheds Light on the Mechanism and Suggests the "Basic Module" for Dimeric d(C/U)TPases. *Journal of Molecular Biology* **342**, 1583–1597 (2004).

27. C. S. Mota, A. M. D. Gonçalves, D. de Sanctis, Deinococcus radiodurans DR2231 is a two-metal-ion mechanism hydrolase with exclusive activity on dUTP. *The FEBS Journal* **283**, 4274–4290 (2016).

28. E. E. Kim, H. W. Wyckoff, Reaction mechanism of alkaline phosphatase based on crystal structures: Two-metal ion catalysis. *Journal of Molecular Biology* **218**, 449–464 (1991).

29. L. Holm, Benchmarking fold detection by DaliLite v.5. *Bioinformatics* https:/doi.org/10.1093/bioinformatics/btz536 (December 10, 2019).

30. J. Pei, B.-H. Kim, N. V. Grishin, PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **36**, 2295–2300 (2008).

31. W. Wang, B. W. Poland, R. B. Honzatko, H. J. Fromm, Identification of Arginine Residues in the Putative L-Aspartate Binding Site of Escherichiacoli Adenylosuccinate Synthetase. *J. Biol. Chem.* **270**, 13160–13163 (1995).

32. O. V. Moroz, *et al.*, Dimeric dUTPases, HisE, and MazG belong to a New Superfamily of all-α NTP Pyrophosphohydrolases with Potential "House-cleaning" Functions. *Journal of Molecular Biology* **347**, 243–255 (2005).

33. F. Javid-Majd, D. Yang, T. R. Ioerger, J. C. Sacchettini, The 1.25 Å resolution structure of phosphoribosyl-ATP pyrophosphohydrolase from Mycobacterium tuberculosis. *Acta Cryst D* **64**, 627–635 (2008).

34. S. Lee, *et al.*, Crystal Structure of Escherichia coli MazG, the Regulator of Nutritional Stress Response. *J. Biol. Chem.* **283**, 15232–15240 (2008).

35. D. W. E. Smith, B. N. Ames, Phosphoribosyladenosine Monophosphate, an Intermediate in Histidine Biosynthesis. *J. Biol. Chem.* **240**, 3056–3063 (1965).

36. L. Chiariotti, P. Alifano, M. S. Carlomagno, C. B. Bruni, Nucleotide sequence of the Escherichia coli hisD gene and of the Escherichia coli and Salmonella typhimurium hisIE region. *Mol Gen Genet* **203**, 382–388 (1986).

37. M. Gross, I. Marianovsky, G. Glaser, MazG – a regulator of programmed cell death in Escherichia coli. *Molecular Microbiology* **59**, 590–601 (2006).

38. A. M. D. Gonçalves, D. de Sanctis, S. M. McSweeney, Structural and Functional Insights into DR2231 Protein, the MazG-like Nucleoside Triphosphate Pyrophosphohydrolase from Deinococcus radiodurans. *J. Biol. Chem.* **286**, 30691–30705 (2011).

39. M. Tsai, *et al.*, Substrate and Product Complexes of Escherichia coli Adenylosuccinate Lyase Provide New Insights into the Enzymatic Mechanism. *Journal of Molecular Biology* **370**, 541–554 (2007).

40. , *SnapGene software (Insightful Science; available at snapgene.com)*.

41. L. Sauguet, P. Raia, G. Henneke, M. Delarue, Shared active site architecture between archaeal PolD and multi-subunit RNA polymerases revealed by X-ray crystallography. *Nature Communications* **7**, 12227 (2016).

42. P. Weber, *et al.*, High-Throughput Crystallization Pipeline at the Crystallography Core Facility of the Institut Pasteur. *Molecules* **24**, 4451 (2019).

43. P. Legrand, *XDS Made Easier (2017) GitHub repository*.

44. D. Liebschner, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst D* **75**, 861–877 (2019).

45. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Cryst D* **66**, 486–501 (2010).

46. E. Krissinel, K. Henrick, Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* **372**, 774–797 (2007).

47. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **42**, W320–W324 (2014).

48. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).

49. E. F. Pettersen, *et al.*, UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

50. , *The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.*

Materials and methods

S-2L genome annotation and identification of related phages

The genomic sequence of cyanophage S-2L was provided by P.-A. K. Potential ORFs were identified using ORFFinder (11) (>150 nt, genetic code 11). Targeted ORFs were assessed for known homologous proteins using BLAST and MOTIF (12). Representation of the S-2L genome was made with SnapGene Viewer (40). Phages related to S-2L through the genes of PurZ and DatZ were found by homology searches using NCBI BLAST (11) and TARA BLAST (19) services.

Protein expression and purification

Synthetic genes for expressed proteins were optimized for *E. coli* and synthesized using ThermoFisher's GeneArt service. Genes were cloned into modified RSF1-Duet expression vector with a TEV-cleavable N-terminal 14-histidine tag (41) using New England Biolabs and Anza (Thermo Fisher Scientific) enzymes. *E. coli* BL21-CodonPlus (DE3)-RIPL cells (Agilent) were separately transformed with engineered plasmids. Bacteria were cultivated at 37°C in LB medium with appropriate antibiotic selection (kanamycin and chloramphenicol), and induced at OD=0.6-1.0 with 0.5 mM IPTG. After incubation overnight at 20°C, cells were harvested and homogenized in suspension buffer: 50 mM Tris-HCl pH 8, 400 mM NaCl, 5 mM imidazole. After sonication and centrifugation of bacterial debris, proteins of interest were isolated from corresponding lysate supernatants by purification on Ni-NTA column (same buffer with 500 mM imidazole). Histidine tags were removed from the proteins by incubation with his-tagged TEV enzyme overnight. After removing TEV on Ni-NTA column, proteins were further purified on Superdex 200 10/300 column with 25 mM Tris-HCl pH 8, 300 mM NaCl. All purification columns were from Life Sciences. Protein purity was assessed on an SDS gel (BioRad). The enzymes were concentrated to 10-19.5 mg ml^{-1} with Amicon Ultra 10k and 30k MWCO centrifugal filters (Merck) and stored directly at -80°C, with no glycerol added.

Nucleotide HPLC analysis

30 µM (1.25 mg ml^{-1}) of S-2L PurZ or 4.2 µM (0.2 mg ml^{-1}) of *E. coli* AdSS was incubated at 37°C for 1h (if not stated otherwise) with 2 mM of respective nucleotides and 10 mM of aspartate, in a buffer containing 50 mM Tris pH 7.5 and 5 mM $MgSO_4$. For S-2L MazZ, 10 µM of the enzyme was incubated at 37°C for 15 min with 100 µM of respective nucleotides, in a buffer containing 50 mM Tris pH 7.5 and 5 mM $MgCl_2$. Reaction products were separated from the protein using 10 000 MWCO Vivaspin-500 centrifugal concentrators and stored in -20°C. Products and standards were assayed separately, using around 40 nmol of each for anion-exchange HPLC on DNA-PAC100 (4x50 mm) column (Thermo Fisher Scientific). After equilibration with 150 µl of a suspension buffer (25 mM Tris-HCl pH 8, 0.5% acetonitrile), nucleotides were injected on the column and eluted with 3 min of isocratic flow of the suspension buffer followed by a linear gradient of 0-200 mM NH_4Cl over 12 min (1ml min^{-1}). Eluted nucleotides were detected by absorbance at 260 nm. High-purity nucleotides and chemicals were bought from Sigma Aldrich, and HPLC-quality acetonitrile was from Serva.

Crystallography and structural analysis

All crystallization conditions were screened using the sitting drop technique on an automated crystallography platform (42) and were reproduced manually using the hanging drop method with ratios of protein to well solution ranging from 1:2 to 2:1. PurZ was screened at 19 mg ml^{-1} with a molar excess of 1.2 of dGMP in 4°C. Capped thick rods grew over several days in 15% Tacsimate and 2% w/v PEG 3350 buffered with 100 mM HEPES pH 7, and did not appear in the absence of dGMP. MazZ was screened at 14.7 mg ml^{-1} with a molar excess of 1.2 of dGTP at 18°C. Big bundles of thin needles grew over a week in 200 mM Li_2SO_4 and 1.26 M $(NH_4)_2SO_4$ buffered with 100 mM Tris pH 8.5. PurZ crystals were soaked for 15 min in a solution containing 30% glycerol, 50% dATP solution (100 mM) and 20% crystallization buffer; MazZ crystals were soaked for several seconds with 30% glycerol and 70% crystallization buffer, with added 30 mM $MnCl_2$. All crystals were then frozen in liquid nitrogen. Crystallographic data was collected at the Soleil synchrotron in France (beamlines PX1 and PX2),

processed with XDSME (43) pipeline and refined in Phenix (44). The structure of S-2L PurZ was solved by molecular refinement with ϕVC8 PurZ model (PDB ID: 6FM1). The structure of MazZ was solved using anomalous signal from bound Mn^{2+} ions that guided automatic model-building in Phenix' AutoSol, with final manual reconstruction steps using Coot (45). PurZ quaternary structure analysis was performed using PISA (46).

Structure and sequence alignments, phylogeny

Structures homologous to PurZ and MazZ available in PDB were identified using Dali server (29). Dali was further used for pairwise RMSD determination. The sequences were aligned in PROMALS3D (30) using structural data supplemented by full protein sequences. Graphical multialignment was prepared with ESPript 3 (47). Maximum-likelihood phylogenetic trees were prepared with MEGA X software (48) with default parameters, taking 100 bootstrap replications. All protein structures were visualised with Chimera (49) and Pymol (50).

Figures

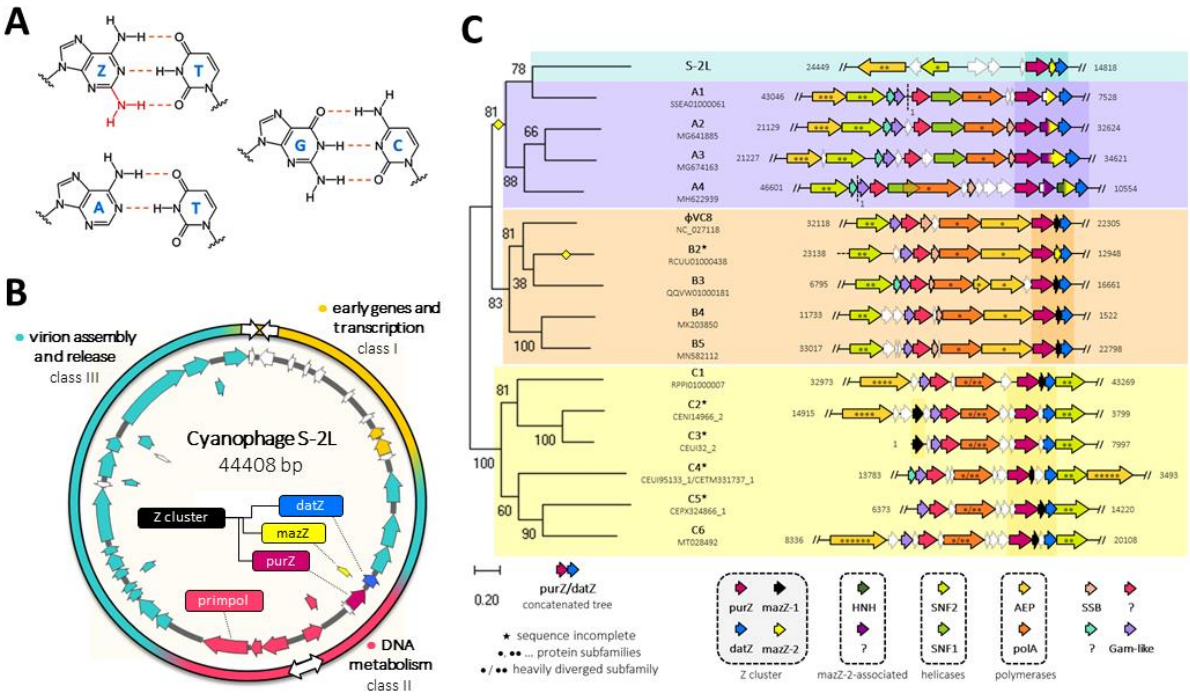


Figure 1. Characteristics of S-2L genome and its closest relatives with Z-cluster. A. Natural basepairs with their Watson-Crick hydrogen bonds. A:T and G:C basepairs are almost universal; the only exception is S-2L's ZTGC-DNA, composed of Z:T and G:C basepairs. **B.** Map of S-2L circular genome. Arrows in the inner ring symbolize genes on the DNA molecule. Arrows on the outer ring delimit portions of the DNA with the same directionality of transcripts. The colouring refers to division into three parts with distinct functionality: early genes (orange, class I), middle ones (magenta, class II), and late ones (cyan, class III). White genes have no identified function. Amid class II genes, involved in DNA metabolism, are DNA Primase-Polymerase (PrimPol) and the cluster of genes involved in dZTP synthesis and dATP removal – the Z-cluster composed of *datZ*, *mazZ* and *purZ*. **C.** Genomic map of most important replication genes in all phages with close *datZ* and *purZ* homologues, and their phylogeny. Due to occasional inconsistency and sometimes erroneous naming, phages other than S-2L and ϕVC8 are described by a code and their NCBI/Tara reference number; incomplete phage sequences are starred. The co-conservation of *datZ*, *purZ* and one of two possible *mazZ* variants is made apparent (darker background). Yellow dots on the phylogenetic tree show probable events of *mazZ-1* to *mazZ-2* gene exchange. The phages can be presently divided into four clades (colours in background) with distinct organisation and variants of replication-related genes. The names of these genes, identified by their colours, are shown below; dots inside the arrows stand for specific gene variants. Cyanophage S-2L is, until now, the only representative of its clade.

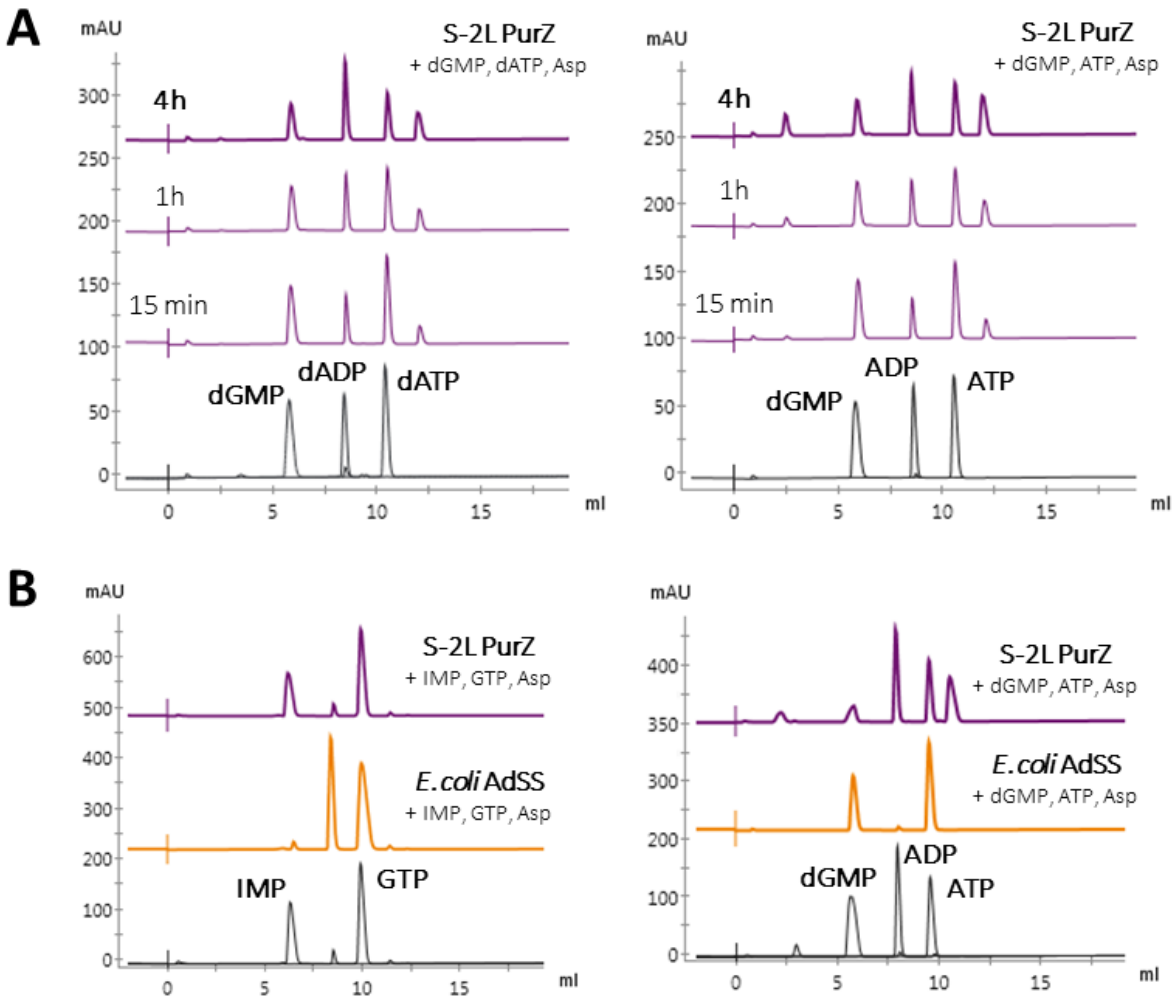


Figure 2. Catalytic properties of S-2L PurZ investigated by HPLC. Reactants of PurZ catalysis are visualised on absorbance chromatograms (purple). Pure compounds were injected in a last run (grey) and are indicated with black labels. **A.** Products of PurZ catalysis with dGMP, Asp and dATP/ATP (left and right panel respectively), taken at three consecutive time steps. The enzyme shows no discrimination between the two adenosine triphosphate variants. **B.** Comparison of the expected activity between PurZ and a typical PurA family representative from *E. coli*, AdSS (orange). In the reaction time of 15 min, AdSS rapidly transforms IMP, GTP and Asp mixture into GDP and sIMP, whereas PurZ stays inactive even at higher concentration (left panel). Inversely, PurZ catalyses the reaction from dGMP, ATP and Asp to ADP and sdGMP, contrary to AdSS that does not recognise these substrates (right panel). Although the sIMP peak is confounded with the GTP one, it gives a noticeable shift in 260/280 nm absorbance ratio.

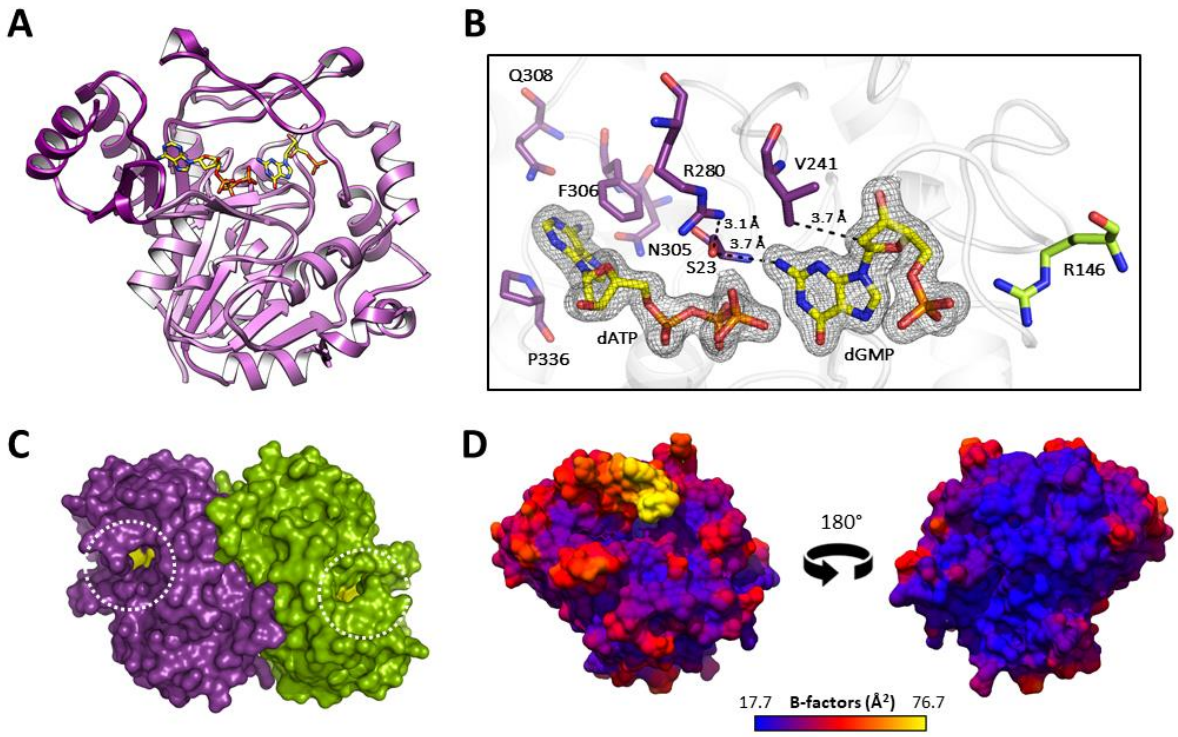


Figure 3. Structure of S-2L PurZ deoxyguanidylosuccinate synthase with ligands dGMP and dATP. A. Ribbon representation of a PurZ monomer in a light purple-dark purple gradient, with dGMP and dATP shown in stick (yellow). **B.** Catalytic pocket of PurZ with the experimental electron density contoured around the reactants at 1 sigma (black mesh). Surrounding residues (purple) are defined in the text; R146 from the second protein subunit (lime) is stabilizing the phosphate of dGMP in the first subunit's catalytic pocket. **C.** PurZ dimer, in surface representation. White dotted circles point to the opposite catalytic cavities. **D.** Surface representation of PurZ coloured using experimental B-factors with a scale-bar at the bottom. The yellow loop above the catalytic cleft (left) define the flexible aspartate loop. The interface between the dimer (right) is particularly rigid, suggesting a constitutive dimeric form of PurZ.

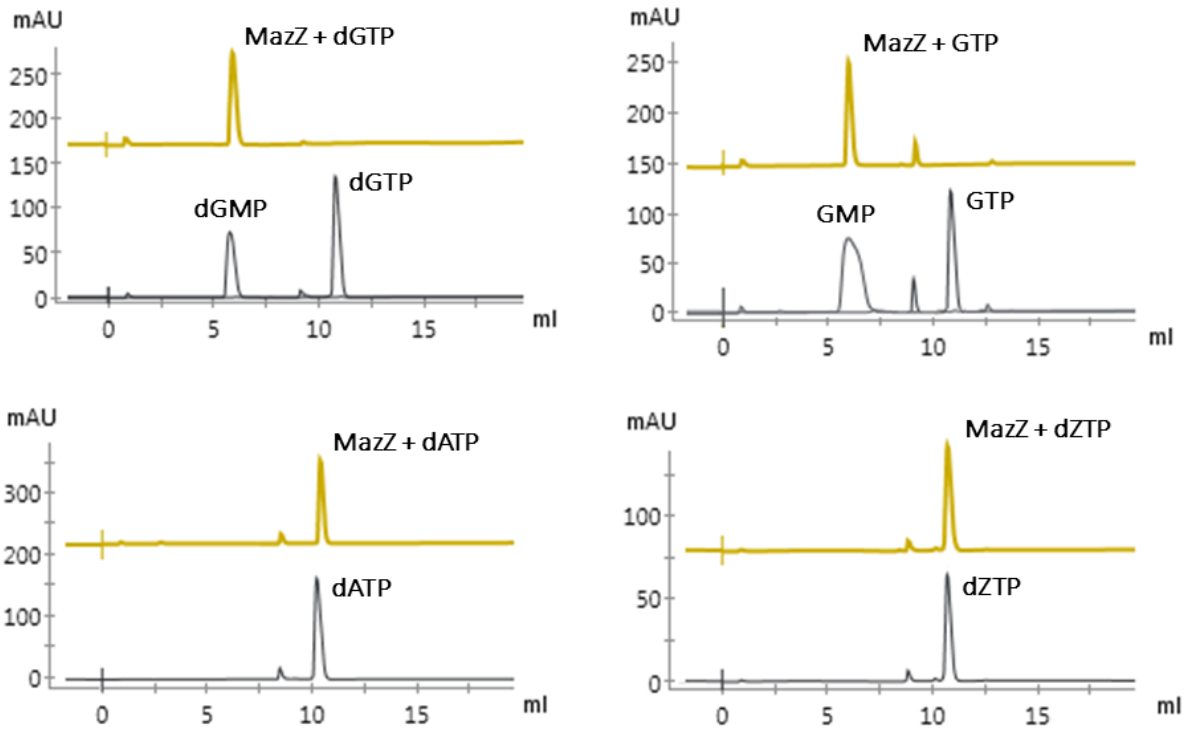


Figure 4. HPLC analysis of S-2L MazZ purine triphosphate specificity and dephosphorylation products. Nucleotide standards are in black, the products eluted after incubation of the corresponding triphosphates with MazZ are in gold. Each sample was eluted separately, after an injection of 40 nmol. The enzyme is selective towards dGTP and GTP, removing their two terminal β- and γ-phosphates.

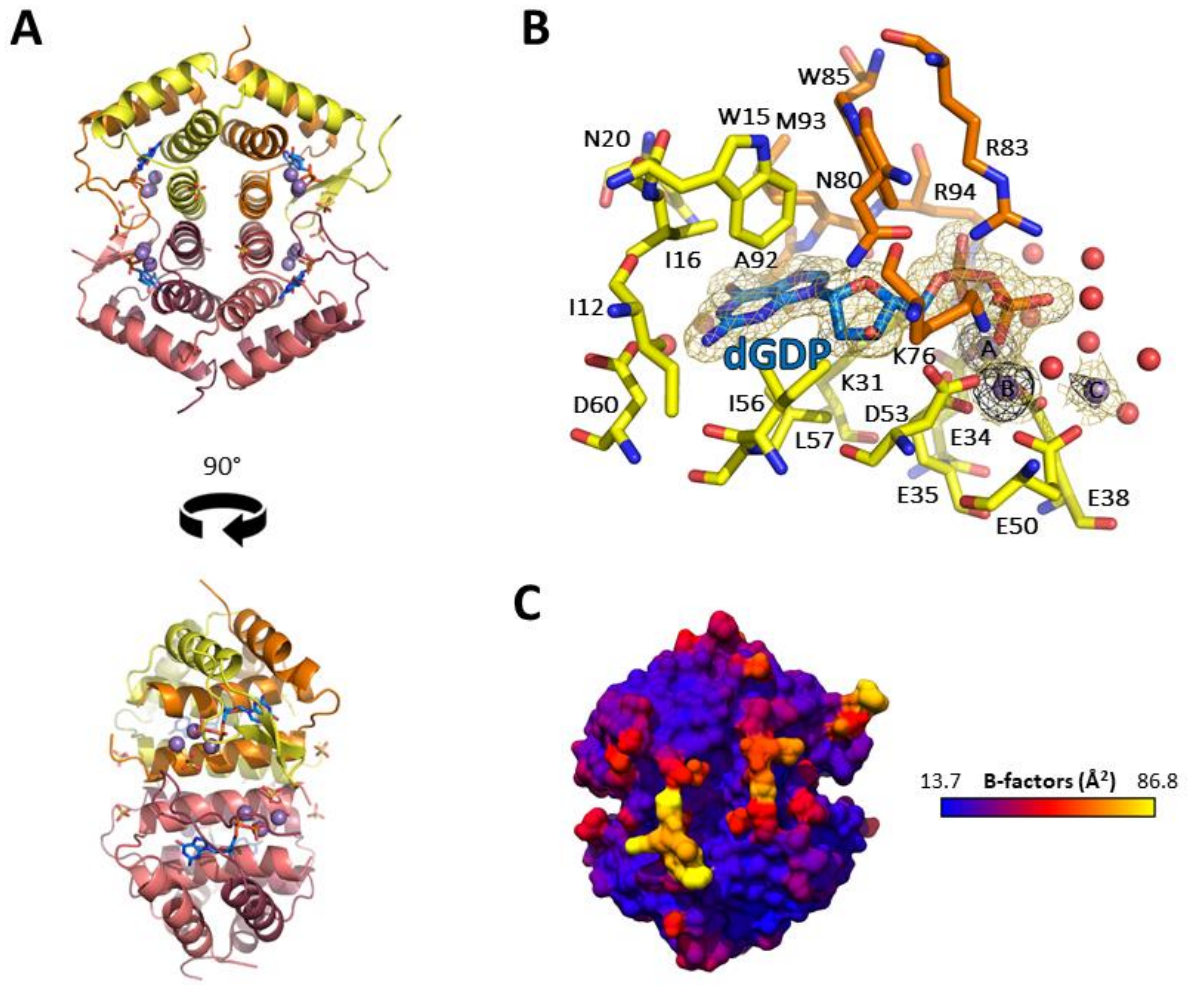


Figure 5. Structure of S-2L MazZ with bound dGDP, Mn^{2+} and SO_4^{2-} ions. A. A tetramer of MazZ, that constitutes the crystallographic asymmetric unit. Four chains (with different colors) form a dimer of tight dimers; each of the four catalytic pockets is created from the two chains of a tight dimer. **B.** Close-up on the catalytic pocket in stick representation, with hydrogen atoms omitted for clarity. The product of dGTP dephosphorylation identified as dGDP in the crystal is completely surrounded by residues of two MazZ chains (yellow and orange). Three hexa-coordinated Mn^{2+} ions (lilac spheres), designated A, B and C, are bound to negatively charged protein residues, deoxynucleotide phosphates and water molecules (red spheres). The 2Fo-Fc electron density around the ligands is contoured at 1 sigma (yellow mesh); the anomalous signal attesting for the presence of Mn^{2+} ions is contoured at 3 sigmas (black mesh). **C.** Surface representation of MazZ coloured using experimental B-factors with a scale-bar to the right. The whole tetramer is rigid, except for the N- and C-termini and the solvent-exposed D43-H46 flexible loop, fully modelled only for the chain A.

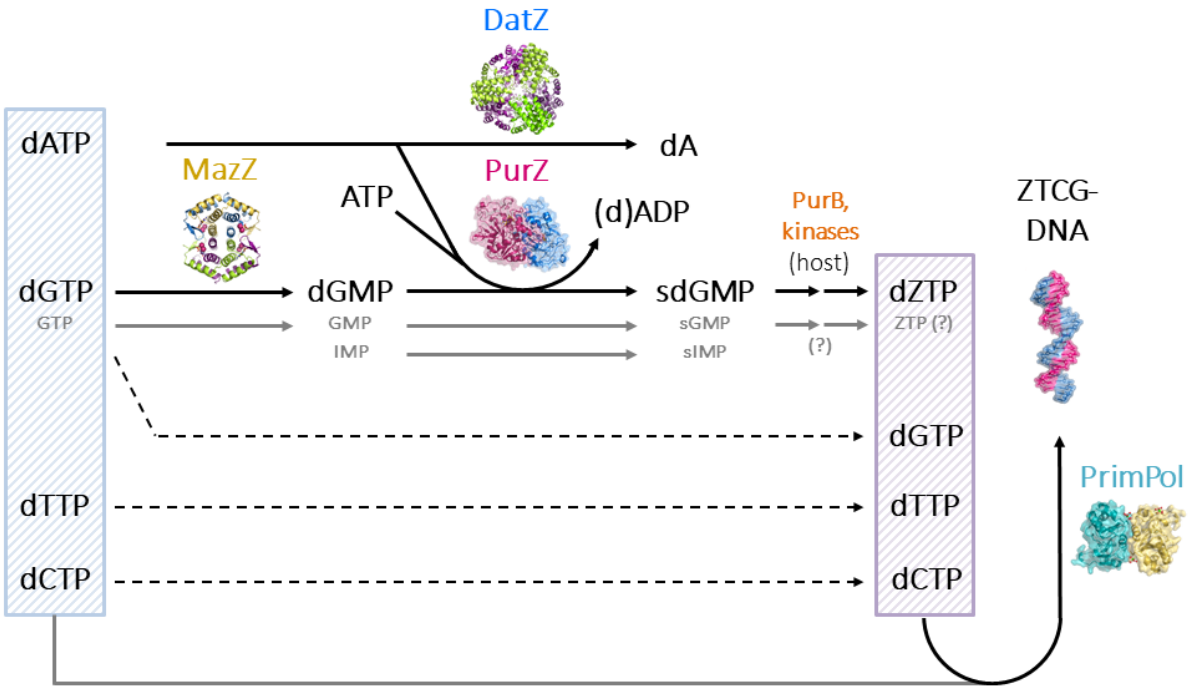
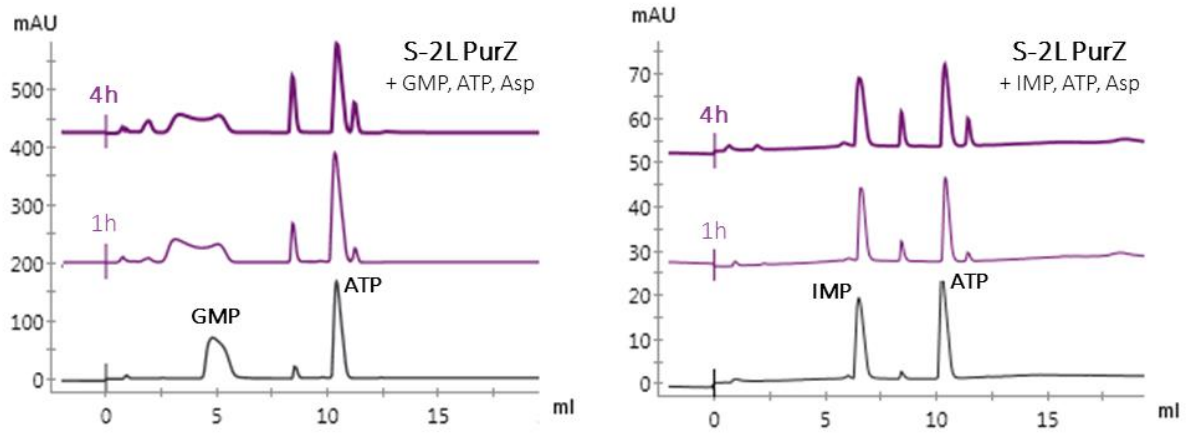
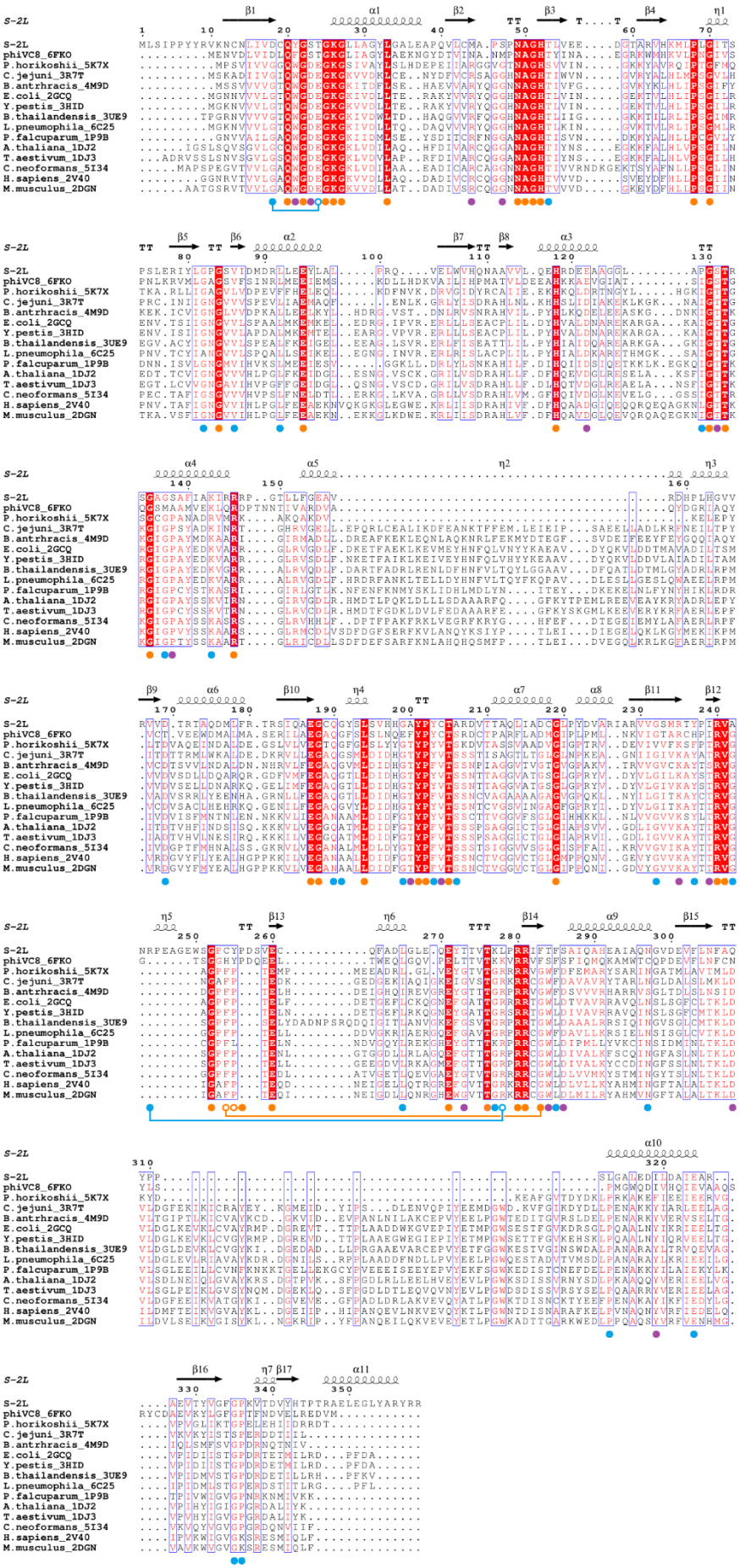


Figure 6. Cyanophage S-2L metabolic pathway of dNTP pool modification and replication. The original pool of nucleotides is in blue stripped background, the modified one is in purple. Structures of S-2L proteins catalysing the reactions solved in this and a previous work (Czernecki *et al*., submitted) are shown next to the corresponding arrows. Thin, dashed arrows stand for no modification. Arrows in grey show possible, but low-yield or side reactions. Host enzymes, in orange, finalise the dZTP pathway (Sleiman *et al*., submitted). Interrogation points indicate possible continuation of the pathway for ribonucleotides, predicted by structural analysis of their homologues in *E. coli*.

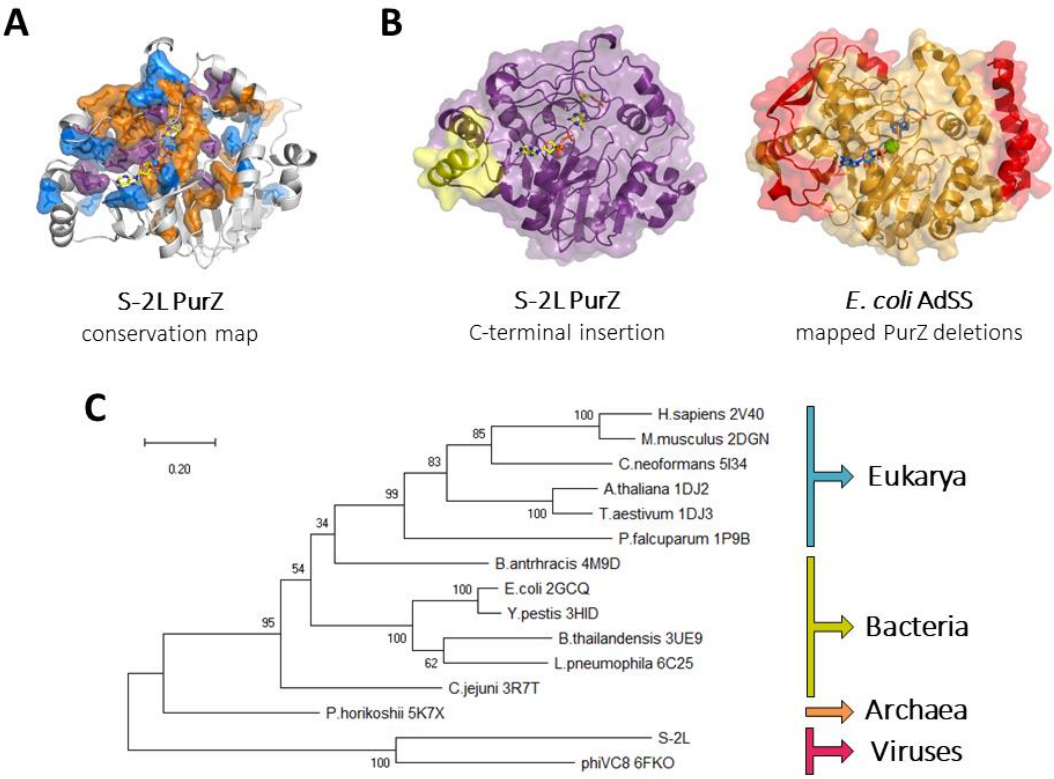
Supplementary Figures



Supplementary Figure S1. Catalytic activity of PurZ with ATP, Asp and GMP or IMP. The detection for the IMP reaction is shown exceptionally at 280 nm due to the low extinction coefficient for IMP and sIMP at 260 nm. ADP and sGMP/sIMP accumulate over longer and much longer periods than for sdGMP, respectively.


S-2L
S-2L
phiVC8_6FKO
P.horikoshii_5K7X
C.jejuni_3R7T
B.antrhracis_4M9D
E.coli_2GCQ
Y.pestis_3HID
B.thailandensis_3UE9
L.pneumophila_6C25
P.falcuparum_1P9B
A.thaliana_1DJ2
T.aestivum_1DJ3
C.neoformans_5I34
H.sapiens_2V40
M.musculus_2DGN

Supplementary figure S2. Structural multi-alignment between S-2L PurZ and all 14 other unique adenylosuccinate/deoxyguanidylosuccinate synthetase crystal structures available in PDB. Organism names and PDB codes are indicated on the left. Circles below the alignment mark positions of residues of interest, divided into three categories: residues strictly conserved across purA family (orange); residues conserved in the usual AdSS enzymes but not in phages (purple); more loosely conserved residues with two similar variants or with only occasional mutations (blue). Empty circles highlight highly conserved residues with shifted backbone position with respect to S-2L's ones (connected full circles), but with similar superposed functional groups. Occasional unstructured and unbuilt regions in the middle were supplemented using the sequence information alone.



Supplementary Figure S3. Visualisations of relationships between S-2L PurZ and homologous proteins. A. Conserved residues from Figure Fig. S2 are mapped onto PurZ structure, using the same colour code. **B.** Visualisation of a S-2L-specific C-terminal insertion in form of a alpha-helix on PurZ structure (yellow, left) and archaeo-viral deletions of two large segments, shown on *E.coli* AdSS (red, right). **C.** Non-rooted maximum-likelihood phylogenetic tree of AdSS representatives using the structural alignment from Fig. S2. Enzymes are divided into four clades: eukaryotic, bacterial, archaeal and viral, the two last ones sharing a recent ancestor. The reference distance corresponds to an average 0.2 substitution per site. The topology of the bootstrap consensus tree is identical, supporting the result presented here.

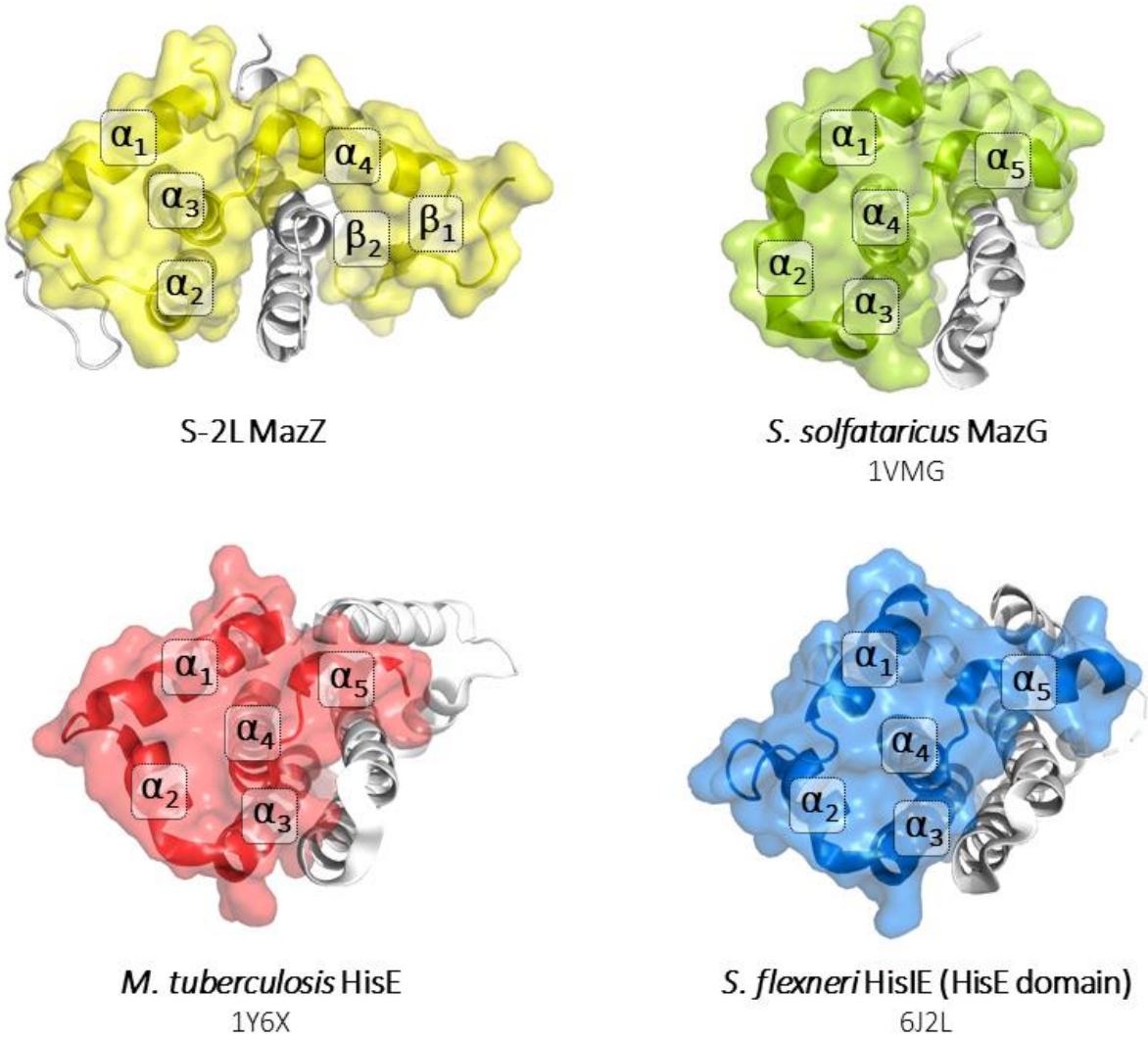


Figure S4. Identical fold shared by bacterial MazG, HisE and S-2L MazZ proteins. The tight dimer part is shown for four exemplary structures, viewed from the same perspective. For each enzyme both chains are shown in ribbon representation: one chain is additionally coloured and its surface traced. For the coloured chain, the secondary structure elements are numbered. Below each image, the organism of origin, protein name and the PDB code are Indicated.

Protein structure	**PurZ + dGMP, dATP**	**MazZ + dGDP, Mn^{2+}, SO_4^{2-}**
Cell parameters		
Space group	P 62 2 2	P 21 21 21
a, *b*, *c* (Å)	108.18, 108.18, 142.33	53.59, 91.37, 114.28
α, *β*, *γ* (°)	90.0, 90.0, 120.0	90.0, 90.0, 90.0
Data statistics		
Resolution (Å)	44.50 - 1.70 (1.74 - 1.70)	48.52 - 1.43 (1.47 - 1.43)
Wavelength (Å)	0.980130	1.127129
Rmerge (%)	8.1 (266.1)	8.9 (193.8)
Completeness (%)	99.8 (98.0)	99.9 (98.3)
Multiplicity	39.6 (39.4)	13.1 (10.9)
$I/\sigma(I)$	34.1 (1.7)	14.1 (1.1)
$CC_{1/2}$	1.000 (0.694)	0.999 (0.579)
Refinement		
Resolution (Å)	42.33 - 1.70	40.45 - 1.43
Unique reflections	54,792	104,269
R_{work}/R_{free} (%)	15.86/17.63	16.09/16.49
No. of non-hydrogen atoms		
Protein	3149	3594
Ligand	53	168
Metal ions	0	12
Water	394	492
Hydrogens added	No	Yes
Protein geometry		
Bond lengths (Å)	0.010	0.008
Bond angles (°)	1.00	0.88
Ramachandran favored/outliers (%)	96.84/0.00	98.92/0.00
Rotamers favored/poor (%)	97.54/0.00	97.00/0.00
B-factors ($Å^2$)		
Type	Anisotropic	Anisotropic
Protein	29.98	25.19
Ligand	28.08	25.89
Metal ions	-	21.12
Water	43.38	41.36

Table 1. Diffraction data collection and Model Refinement statistics. Numbers in parenthesis refer to the highest-resolution shell.

III. Structure-function of the Z-selective DNA polymerase from bacteriophage ϕVC8

1. Introduction

The strategy adopted by S-2L to incorporate Z instead of A in front of a T in its genome, based on a very efficient dATPase, is only one among several possible ones. Indeed, it was found by our collaborators – P.-A. Kaminski (*Institut Pasteur*) and P. Marlière (*Institut de Biologie Systémique et Synthétique* – iSSB, Evry) and their colleagues – that some phages that do contain Z in their genomes have a specific family A DNA polymerase (PolA) with a strong specificity of Z vs A incorporation during DNA replication (hereafter referred to as PolZ).

Here, I describe, both functionally and structurally, this novel functional category of DNA polymerases from bacterial viruses in which 2- aminoadenine (Z) fully replaces adenine in the genome. PolZ discriminates against A and for Z in nucleotide incorporation templated by T. The crystal structure of PolZ from a *Vibrio cholerae* phage ϕVC8 reveals the typical fold of the PolA family of DNA polymerases. The polymerase active site does not show any major mutations compared to the consensus Motifs A, B, C, present in all PolA. However, the enzyme is found in two conformations that open and close an insertion in the exonuclease active site – an unusual PolA feature seen here for the first time – suggesting an important role for the PolZ proofreading activity. Therefore, the enzyme possibly acquired part of its specificity by systematically sending any newly-formed A:T base pair to the highly active exonuclease site, contrary to a Z:T base pair. Thus, due to the lesser stability of the former base pair, adenine would be removed in the proofreading stage, while Z would be translocated to the next position.

Crystallization of vibriophage ϕVC8 DNA polymerase from family A (PolZ), bearing specificity towards dZTP, was the structural part of a shared project that extended outside the laboratory of M. Delarue. The main idea of this project was that if there exists a DNA polymerase (in this case from the PolA family) in a phage whose genome contains Z and no A, then it should be worth studying at the structural level how the DNA polymerase selects a Z in front of a T and rejects the A. Functional tests indeed showed that it is the case, namely that this DNA polymerase is selective for Z vs A.

The crystallization of this DNA polymerase was the subject of a 2-month internship during my graduate studies at the *Ecole Normale Supérieure*, one year before I returned to the unit for the

doctorate. During this short stay, I managed to purify and crystallize PolZ (Fig. 29). In the interim of my absence, my colleagues optimized PolZ crystals, collected diffraction data (F. Romoli) and solved the enzyme's structure (H. Hu) at 2.69 Å resolution.

Figure 29. First crystal of PolZ, which served as a seed source for crystal optimization through microseeding.

After my return, I refined the structure of PolZ and analyzed it as described below. This structural part will be included in a scientific article describing family A DNA polymerases of ZTGC-DNA bacteriophages, presently in preparation (Pezo *et al.*). I also performed a brief structure-function analysis, generating several mutants and testing their activity, which I include in a following section. Finally, to help understand the variability in the PolA family, I provide in an appendix to this chapter an attempt to re-classify of all extant PolA sequences found in databases, about 33,000 sequences. It highlights the special position of the PolZ sequences among all PolAs and describes a novel actinobacterial PolA subfamily, Pol I D.

2. Crystallographic methods

To characterize the structural determinants of PolZ function, the crystal structure of ϕVC8enzyme (produced from *E. coli* as N-terminal His-tagged version) was solved by X-ray crystallography at 2.79 Å resolution, with good statistics for data collection and refinement (Table 6).

The crystallization conditions for ϕVC8 PolZ were determined with the sitting drop technique on an automated crystallography and crystallogenesis platform (179) and were reproduced manually with the hanging drop method. The protein was screened at 10 mg ml^{-1} at 18°C. Several small crystals grew during several days in a mother liquor containing 1 mM Hexamine Cobalt and 25% v/v isopropanol (100%) buffered with 100 mM HEPES pH 7. These crystals were optimized by seeding in a solution containing 12.5 mg ml^{-1} of the protein, 300 mM ammonium citrate and 12% PEG 3350. Final crystals were soaked in a solution composed of 70% crystallization buffer and 30% glycerol, before being frozen in liquid nitrogen. Crystallographic data was collected at beamline PX2 in the Soleil synchrotron in Saint-Aubin, France. It was processed with XDSME pipeline (180) and refined in Phenix (181), mixed

with manual reconstruction steps using Coot (182). The structure was solved by the molecular replacement technique using the *E. coli* pol I (Klenow fragment) as a template model. The solution was confirmed by a data set with crystals soaked with a platinum derivative, which, however, could not be used for model building as its phasing power was effective only at medium resolution (4 Å).

The Normal Modes calculations (183) were performed using the NOMAD-Ref web server (184).

Protein structure	ϕVC8 PolZ
Cell parameters	
Space group	P2(1)2(1)2
a, *b*, *c* (Å)	120.16 158.44 79.95
α, *β*, *γ* (°)	90, 90, 90
Data statistics	
Resolution (Å)	48.35 - 2.79 (2.87 - 2.79)
Wavelength (Å)	1.6926
Rmerge (%)	35.2 (261.7)
Completeness (%)	99.4 (92.7)
Multiplicity	26.5 (23.5)
$I/\sigma(I)$	9.7 (1.4)
$CC_{(1/2)}$	0.996 (0.927)
Refinement	
Resolution (Å)	48.03 - 2.797
Unique reflections	38380
R_{work}/R_{free} (%)	18.92/25.82
No. of non-hydrogen atoms	
Protein	9575
Water	134
B-factors (Å^2)	
Protein	11.64
Water	25.26
Protein geometry	
Bond lengths (Å)	0.008
Bond angles (°)	0.95
Ramachandran favored/outliers (%)	95.28/0.00
Rotamers poor/favored (%)	4.72/95.28

Table 6. Crystallographic statistics for the structure of phiVC8 PolZ.

3. Description of ϕVC8 PolZ structure

The enzyme exhibits the typical fold of the PolA family (Fig. 30A), with two separate domains (185):

the 3’-5’ exonuclease domain and the polymerase domain. The last one resembles a right-hand ready to grip DNA with its palm domain carrying the catalytic site, and both thumb and finger domains sensing the DNA substrate. It contains all the characteristic motifs of a PolA DNA polymerase (186) (Fig. 30B) and superimposes very well on E. coli Klenow fragment of DNAP Pol I, both on the pol domain (Fig. 30C) and the exonuclease domain (Fig. 30D).

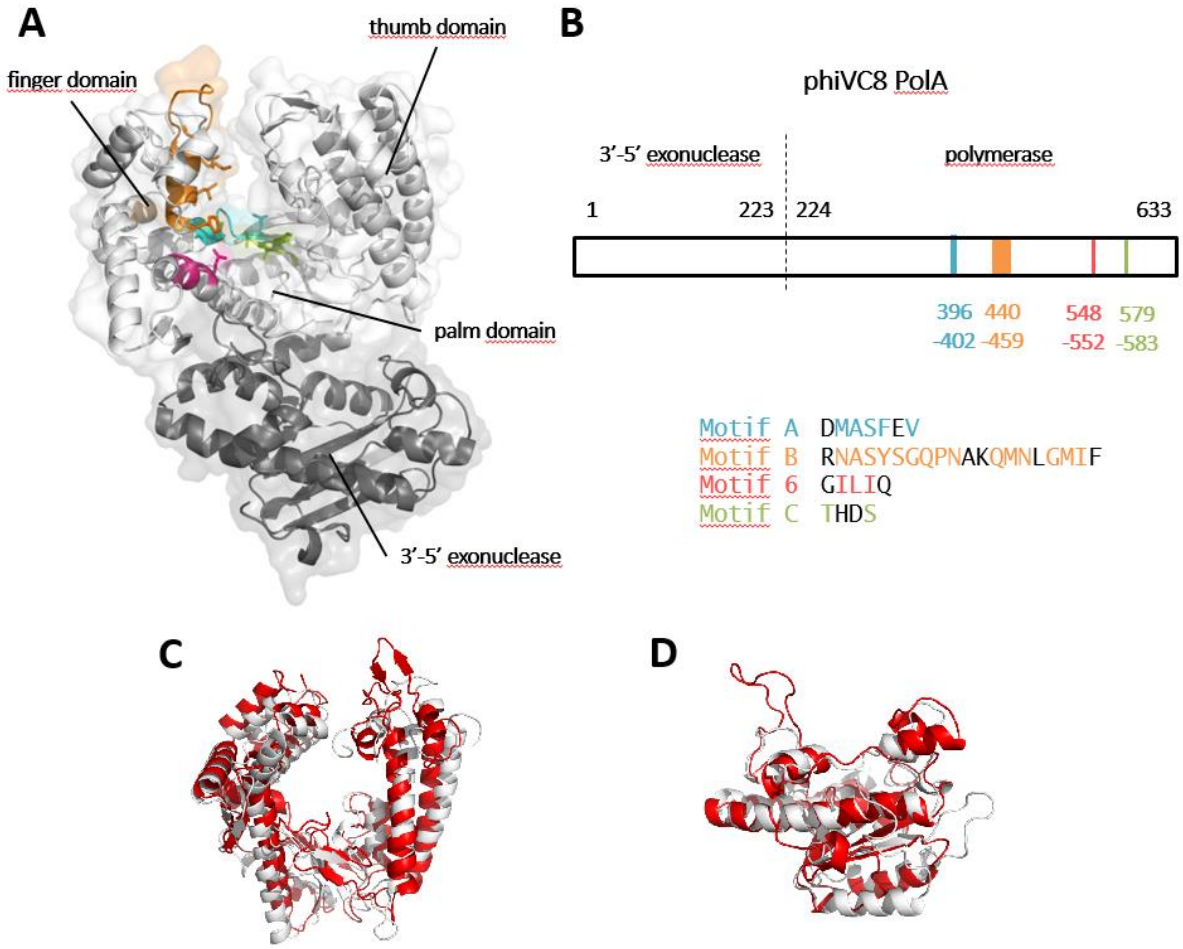


Figure 30. Crystal structure of ϕVC8 PolZ and comparison with *E. coli* Pol I. A. The 3’-5’ exonuclease domain is in dark grey and the polymerase domain in light grey. Conserved sequence motifs with residues crucial for polymerase activity are mapped to the structure and colored: motif A (cyan), motif B - helix O (orange), motif 6 - helix Q (magenta) and motif C (green) mapped onto open structure of ϕVC8 PolZ. B. Position of the conserved motifs on PolZ chain with the same color code (above) and their sequence (below). Conserved residues are in black. C. Superposition of the polymerase domain of *E. coli* Pol I (white) and ϕVC8 PolZ (red). D. Superposition of the exonuclease domain of *E. coli* Pol I (white) and ϕVC8 PolZ (red).

The asymmetric unit in the crystal contains two molecules in different conformations, hereafter referred to as the “coupled-open” and ”coupled-closed” states (Fig. 31A). The two structures display an apparent blocking and unblocking of the DNA binding sites both in the exonuclease and polymerase domains in a highly coupled fashion. Their RMSD is 2.0 Å for the whole structure and 1.8 Å for the polymerase module only. Aside from this peculiarity, the PolZ structure shows several novel insertions with respect to typical PolA enzymes (Fig. 31B), but maintains all standard catalytic residues located in polymerase (Fig. 31C) and exonuclease (Fig. 31D) domains.

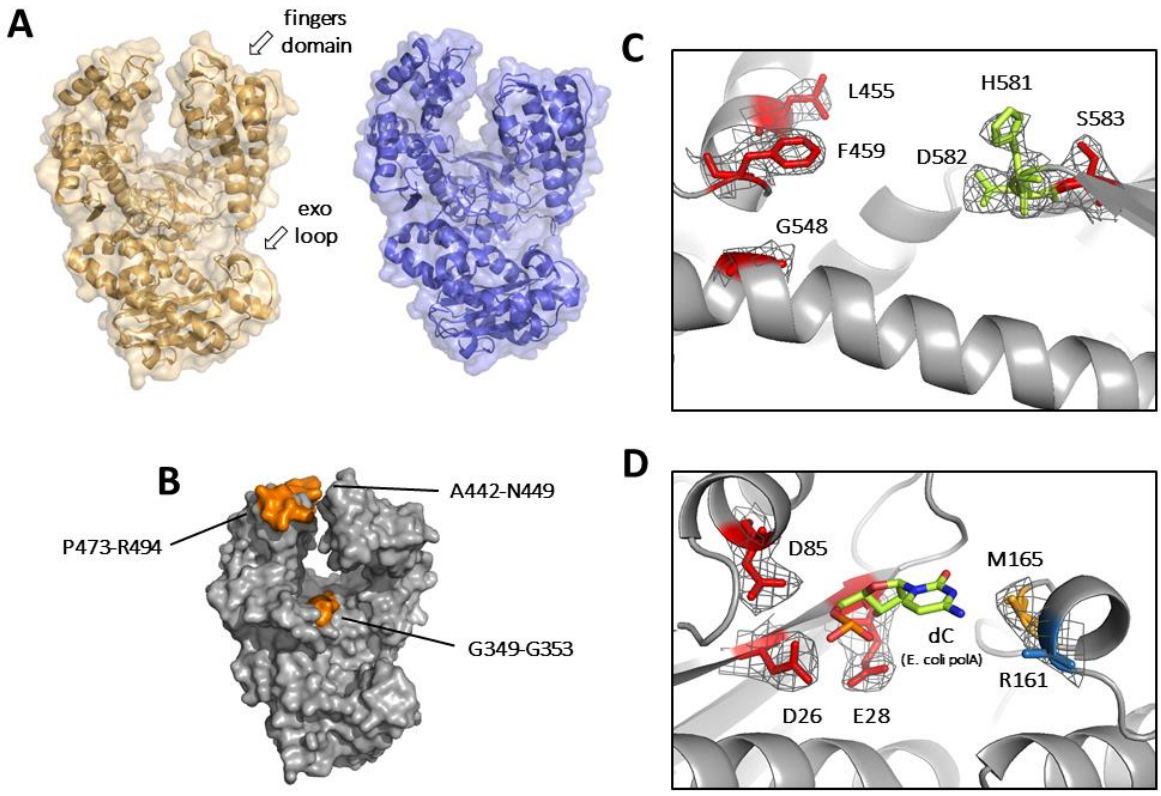


Figure 31. A. Two states of PolZ, captured in the asymmetric unit: coupled-open (left, pale yellow) and coupled-closed state (right, blue). White arrows indicate domains that undergo a large motion: the thumb domain in the polymerase domain and a flexible loop in the exonuclease domain (exo loop). B. Insertions specific to ϕVC8 PolZ are mapped to the open state, in orange. C. Catalytic center of polymerase domain in the open state. Catalytic residues are marked in yellow-green; in red are rare mutations found in ϕVC8 PolZ. The electron density contoured at 1 sigma is shown as a black mesh around the residues of interest. D. Catalytic center of exonuclease domain in the open state. The nucleotide to be excised is modelled after PDB 1KLN. The aspartate catalytic residues (red) are conserved in all PolA. The flexible loop exhibiting the largest movement between the coupled-open and coupled-close states is represented and the position of residues R161 and M165 is indicated: their side-chain (not modelled) would possibly contact specifically the adenine base in the coupled-open state. The density at 1 sigma is shown as a black mesh around the residues of interest.

The highly coupled nature of the conformational transition between the 'coupled-open' and 'coupled-closed' states is well explained by a handful of low-frequency Normal Modes (Fig. 32) derived using a coarse-grained model based on the Elastic Network Model (183).

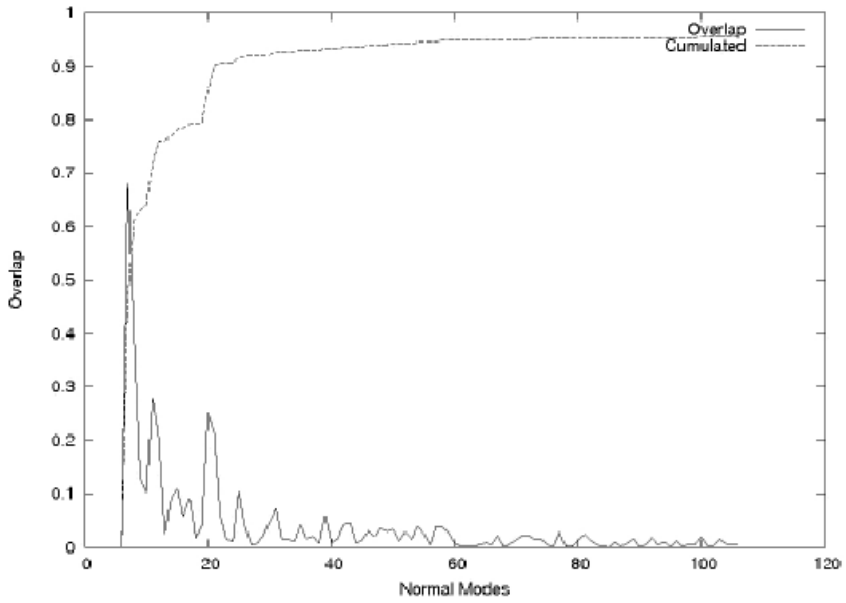


Figure 32. Overlap coefficient between the Normal Modes calculated with the Elastic Network Model (ENM) and the difference vectors between the two coupled-open and coupled-closed forms. The cumulative overlap coefficient is in dotted line. Only a handful of low-frequency normal modes derived from the Model account for 80-90% of the transition between the two forms seen in the asymmetric unit.

The transition between these two conformations involves the simultaneous movement of two smaller domains moving mostly as rigid-bodies: the thumb domain holding the nascent double-stranded DNA (185) and a flexible loop in the exonuclease domain. This is different from the open and closed forms observed in the Klentaq PolA structure (187), which differ mainly in the position of the fingers domain and in the opening and closing of the polymerase module, and no change in the exonuclease active site. The smallest RMSD with the closest structure found using the server Dali is 2.9 Å for the coupled-open form and 3.4 Å for the coupled-closed form, both with PolA from *G. stearothermophilus* (PDB ID: 4NQQ).

Compared to other members of the PolA family, we note the presence of three extended loops, unique to ϕVC8 PolZ (Fig. 31B): one loop is located in the Motif 2 between strands 7 and 8 in the palm domain (residues G349-G353), another loop is in helix O (Motif B) in the finger domain (residues A442-N449) and the last one is between helices O and P, also in the finger domain (residues P473-

R494), lacking however corresponding electron density in the structure. The Loop G349-G353 is most possibly contacting the template strand in the newly synthesized DNA duplex (Fig. 33); other insertions, stretching around the catalytic site (Fig. 31B), can be involved in direct DNA binding as well.

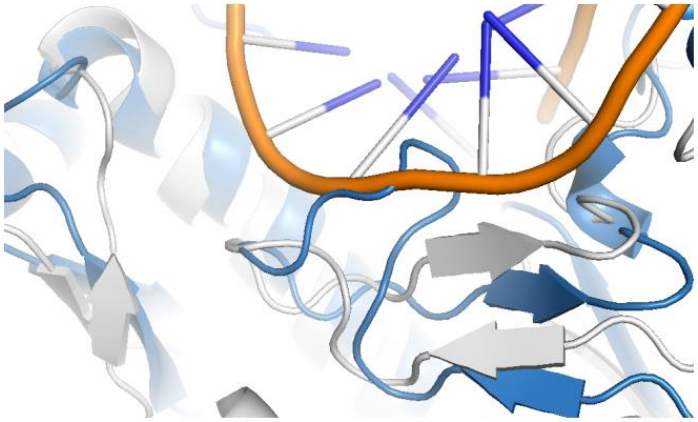


Figure 33. Superposition of *T. aquaticus* Pol I (KlenTaq) bound to dsDNA (white) and ϕVC8 (Pt-soaked, blue) polymerase structures, emphasizing a possible rearrangement of the palm insertion G349-T354 with the DNA in the polymerase catalytic pocket.

All usual PolA motifs were found to be conserved in ϕVC8 PolZ and few departures from the consensus sequences were observed, especially the insertion in motif B on thehelix O (Fig. 34).

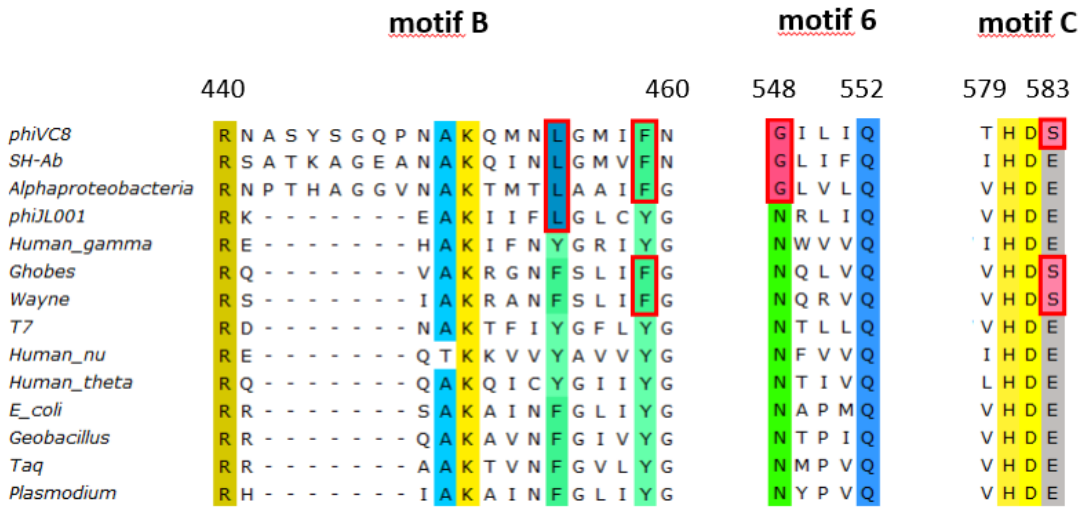


Figure 34. Crucial motifs in ϕVC8 PolZ (top), DNA polymerases of related phages (middle) and typical PolA enzymes with resolved crystallographic structure(bottom).

It is worth noting that the loop in the exonuclease domain that undergoes the opening and closing motion contains two residues – R161 and M165 – that occupy a special position in the catalytic pocket of the domain (Fig. 35A). Although their side chains could not be modelled due to high flexibility, their C_β atoms are in the immediate vicinity of a nucleotide modelled from the *E. coli* Pol I structure (PDB ID: 1KLN). Conceivably, these two residues could confer specificity in the removal of a deoxyadenosine incorporated in response to thymine by the polymerase domain of DNA (Fig. 35B). This issue was explored in catalytic tests on ϕVC8 PolZ exonuclease mutants in the following section.

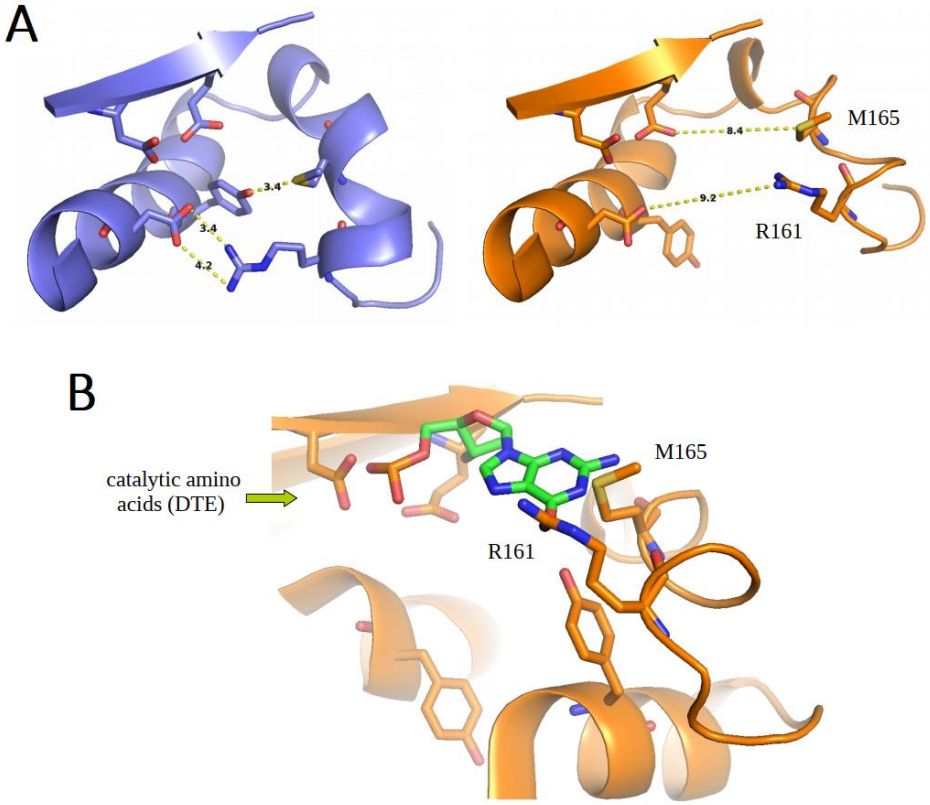


Figure 35. ϕVC8 PolZ exonuclease site (Pt-soaked). A. Movement of the helix containing R161 and M165 in the couple-closed (left) and couple-open state (right). B. Coupled-open state of ϕVC8 PolZ with a model of dGMP in the exonuclease catalytic pocket, modelled from *E. coli* Pol I (PDB ID: 1D8Y).

4. ϕVC8 PolZ functional assays

In addition to the catalytic tests made by our associates at iSSB, Evry, who discovered this enzyme together with P.-A. Kaminski, I performed several catalytic assays on family A polymerases (*E. coli* Pol I Klenow fragment and PolZ) in order to examine their selectivity towards dATP and dZTP. I also

tested mutants of potentially important residues in the exonuclease active site of PolZ, identified through careful studies of the structure. All these assays were made with 3 µM of polythymine (dT_{24}) template, primed by 1 µM of FAM-marked primer complementary to template upstream sequence and with various concentration of dATP or dZTP. Reactions were buffered in 5 mM $MgCl_2$ and 20 mM TRIS pH 7 in 100 µl volume. DNA polymerases were added: PolZ to 0.83 µM (6 µg ml^{-1} or 2.5 µg) or *E. coli* Klenow to 5 U. Assays were conducted at 37°C for 5 min. They were blocked, stored and visualized identically to other polymerase assays.

In the first set of assays (Fig. 36) I show that whereas *E. coli* Pol I (Klenow) is unspecific for Z or A and does not degrade the primer substrate noticeably, PolZ has a higher specificity for dZTP and a high 3'-5' exonuclease activity. The same effect, albeit slightly shifted, was seen in presence of 200 mM NaCl.

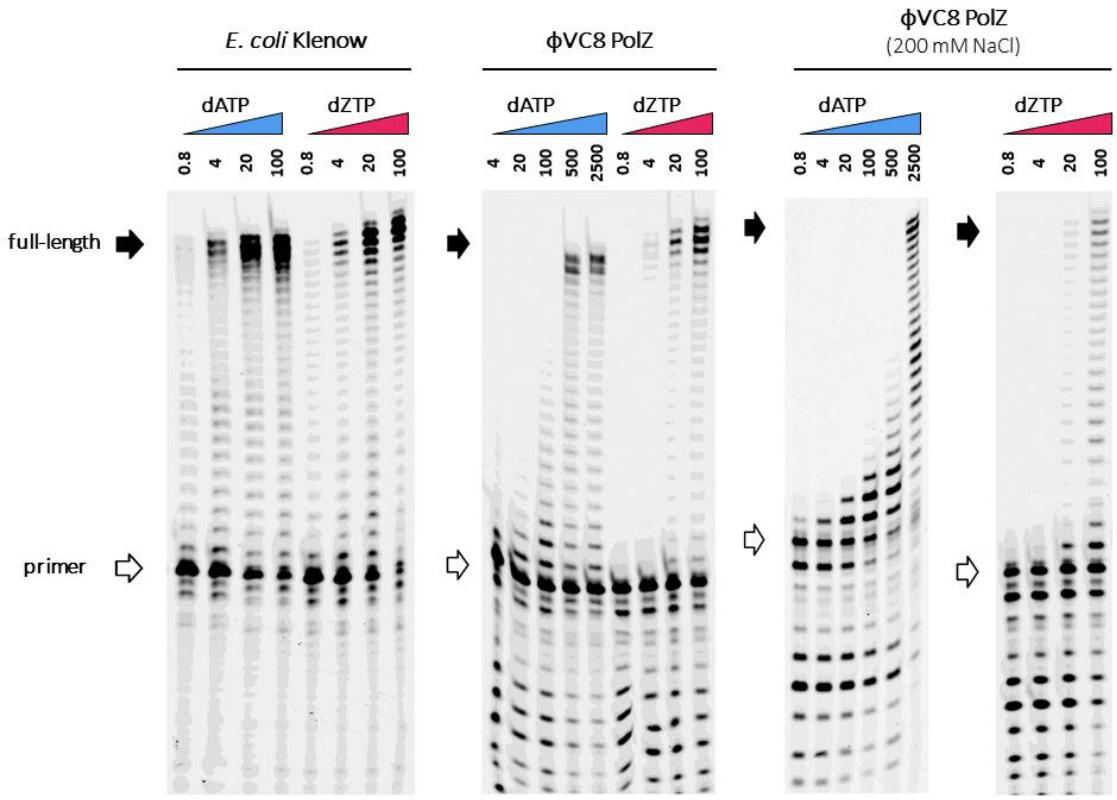


Figure 36. Catalytic assays for *E. coli* Klenow fragment and ϕVC8 PolZ visualized on polyacrylamide gels. The amount of nucleotides used (dATP or dZTP) is indicated above each lane, in µM concentration. The templating substrate is T_{24}.

In the second approach, I mutated residues in PolZ exonuclease domain known for their contribution for the proofreading activity (188), namely D26, E28 and D85 and investigated their enzymatic activity

(Fig. 37). Firstly, native PolZ rapidly digests the primer, in parallel with its simultaneous elongation. The enzyme gradually loses its exonuclease activity with cumulative exonuclease domain mutations, with sharp decrease for the single mutant, enabling full-length polymerization with lower dNTP concentration. Concomitant with this weaker exonuclease activity, we see an increase in dATP incorporation. Triple PolZ mutant shows also lower polymerase activity, as indicated by a higher fraction of non-elongated primers.

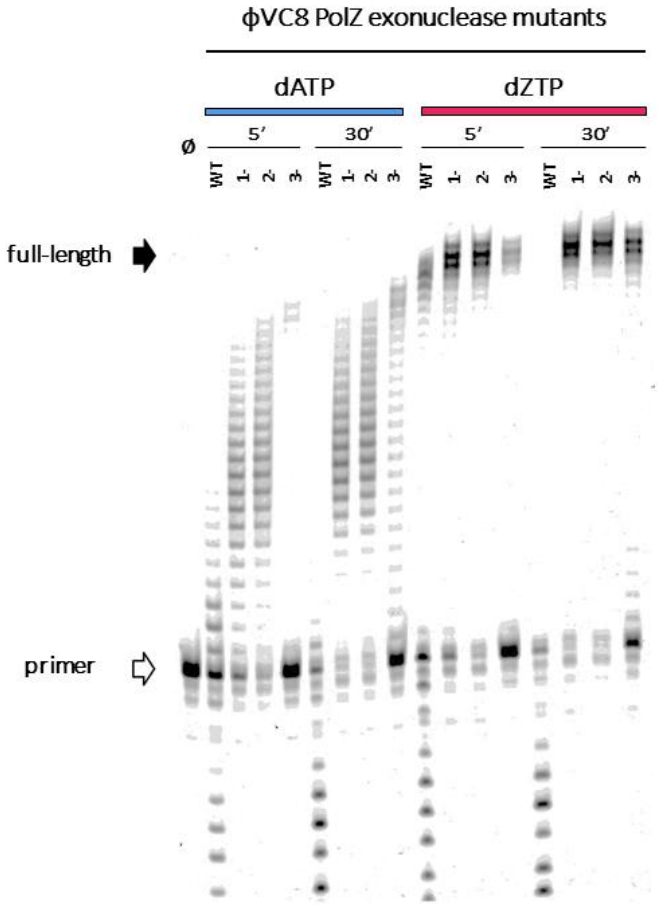


Figure 37. Catalytic assays for ϕVC8 PolZ exonuclease mutants: D85A (1^-), D26A/E28A (2^-), D26A/E28A/D85A (3^-). The amount of nucleotide used (dATP or dZTP) is 100 µM. The assay was done for two incubation times, 5 and 30 min.

Finally, I mutated residues R161 and M165 found to be crucially placed in the exonuclease pocket and conserved between close relatives, and observe lower exonuclease activity of their alanine derivatives (Fig. 38, left). Ultimately, it looks like every impairment of the exonuclease activity allows for a better incorporation of dATP. However, the question arises: does it modify the Z vs A specificity of incorporation?

In response to this problem, we measured the concentration of dATP or dZTP necessary for full-length extension of the primer, both for the wild-type (Fig. 36) and the mutants (Fig. 38). The results, gathered in Table 7, show that the ratio of the base specificity is almost constant for all constructs tested, being roughly 50 times greater for 2-aminoadenine. Therefore, the specificity of PolZ is not modified upon inactivation of the exonuclease.

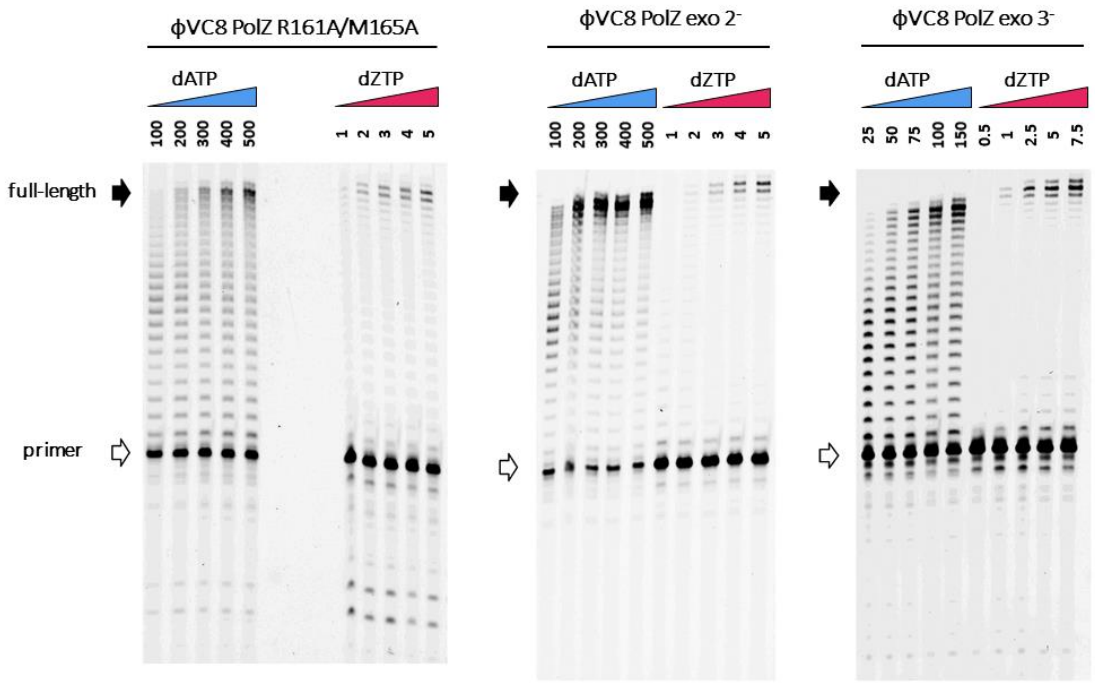


Figure 38. Last catalytic assays for PolZ exonuclease mutants. PolZ R161A/M165A with lower exonuclease activity (left); mutants D26A/E28A (2^-) and D26A/E28A/D85A (3^-) (right).

	E.coli Pol I	PolZ	PolZ 2^-	PolZ 3^-	PolZ RM
dATP	0.8	250 (±100)	100	50	100
dZTP	0.8	4	2	1	1-2

Table 7. Minimal nucleotide concentration in µM necessary for full-length product synthesis in the usual conditions used for the DNA polymerase assays (Fig. 36, 38).

To summarize, a possible interpretation, coherent with all the results obtained, is the following: it is not the exonuclease activity *per se* which confers base specificity, but rather the ability to promote

the transition from the extending mode to the proofreading mode, by sensing the nature of the base pair. To pin-point exactly where in the structure this ability resides is a difficult task. If one looks at the nascent base pair pocket in the polymerase active site, there is seemingly nothing special that would explain the discrimination of a correct (Watson-Crick) nascent base pair, compared to other DNA polymerases of PolA family (*e.g. E. coli* Klenow Pol I or Bst Pol I). Indeed, the volume of a Z:T base pair is exactly the same as for the A:T one. The difference that one should look for is subtle and may have to do with an allosteric network of residues participating in the transition from the elongation (polymerase) mode of the enzyme to the proofreading (exonuclease) one. This communication pathway has baffled researchers in the field for many years, and might actually be different in different families. In the case of ϕVC8 PolZ, , the structure of the ternary complex of the enzyme with both the DNA duplex and the incoming dNTP would be needed to identify residue(s) who might drive this transition, more or less efficiently, depending on the nature of the nascent base pair. So far, our attempts to get crystals of this ternary complex have remained unsuccessful.

PolZ retains its dZTP specificity in all tests, which is high, but not absolute. It is also important to remember that bacterial dATP pool is considerable, at the level of 175-181 µM (189), which is at the threshold of PolZ dATP specificity. It is possible and quite likely that for complete A for Z substitution, the selectivity of PolZ has to be supplemented by the dATPase activity of DatZ, which is co-conserved with the former in all related phages.

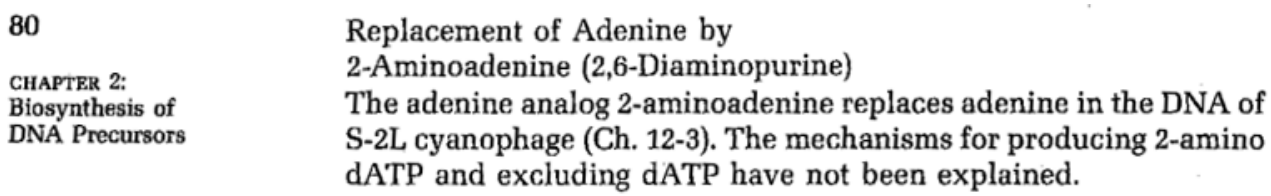
PERSPECTIVES

I. General discussion

1. 2-aminoadenine metabolic pathways

My thesis constitutes an effort to understand the complete metabolism of 2-aminoadenine, a rare natural base substitute for adenine in some phages' DNA. This base may be linked to the origin of life itself, arising from the prebiotic reactions (190) and enabling efficient nonenzymatic replication nucleic acids under specific conditions (191). Its true uniqueness lies however in an additional amino group in respect to the standard analogue that operates at the heart of the Watson-Crick base-pairing scheme. Due to this fundamental change, the source of unusual biophysical properties of ZTGC-DNA, base Z may be regarded as a natural "fifth element".

This fact was promptly recognized at the time of cyanophage S-2L discovery, roughly four decades ago (155). However, at the time of beginning my doctoral research, the first biochemical pathways involving base Z identified in several bacteriophages were only starting to undergo rigorous characterization (Pezo *et al*., in preparation; Sleiman *et al*., submitted). My research, building up on this work, solves a biological problem that laid unexplained for more than 40 years (Fig. 39). Additionally, throughout my studies I acquired a solid experience with viral DNA and RNA polymerases – S-2L PrimPol, ϕVC8 PolZ and SARS-CoV-2 RdRP (see *Annex* section).



80

CHAPTER 2:
Biosynthesis of
DNA Precursors

Replacement of Adenine by
2-Aminoadenine (2,6-Diaminopurine)
The adenine analog 2-aminoadenine replaces adenine in the DNA of S-2L cyanophage (Ch. 12-3). The mechanisms for producing 2-amino dATP and excluding dATP have not been explained.

Figure 39. Extract from A. Kornberg's book *DNA replication* (147), published in 1980.

2. DNA replication: ϕVC8 PolZ and PolA enzymes

Historically speaking, I started my thesis with the study of the structure and function of ϕVC8 DNA polymerase PolZ, an enzyme belonging to the PolA family with a 5'-3' polymerase and 3'-5' exonuclease domain. Being a member of a collaboration focused on explaining selective DNA replication with 2-aminoadenine, my goal was to understand how ϕVC8 PolZ manages to incorporate

preferentially base Z in front of T, inserting A only occasionally (Pezo *et al*., in preparation). PolZ structure was solved at a good resolution, allowing me to describe its unique features, especially in the exonuclease domain. My functional assays on PolZ mutants in key exonuclease positions made me realize that its high proofreading activity was not sufficient to explain the observed specificity.

This discovery made apparent that to truly understand PolZ peculiar functionalities, such unbound (apo) structure is insufficient: the structure of PolZ ternary complex with the DNA duplex and the incoming dNTP was seemingly necessary. It could support the idea that it is the transition between polymerase and exonuclease modes that is sensitive to the nature of a basepair, and specifically, it strength. This activity, linked to polymerase backtracking, well described for DNA-dependent RNA polymerases (192) would result in halting primer extension with weaker A:T bond even with a catalytically inactive exonuclease domain. Such behavior, leading to excision of as much as 7% of new nucleobases under usual conditions for various PolA may be further enhanced by a slowed helicase, as shown for phage T7 (193). It is logical to assume that by its strongly-hybridized nature, ZTGC-DNA has a slow uncoupling rate with standard helicases. Despite having this promising explanation, my extensive crystallization trials towards the obtention of PolZ ternary complex structure proved unsuccessful. Such information, complemented with more holistic assays with other DNA replication actors, like ϕVC8 helicase, may be important to understand the global incorporation mechanism and the basis of its selectivity and could be the subject of future studies.

An important outcome of my research on ϕVC8 PolZ is the new classification of the whole PolA family, described in the *Annex* section. Not only it allowed me to place PolZ and a closely related PolA of ϕJL001 among standard PolA representatives, but also to identify a novel, but highly represented subfamily of PolA of phylum Actinobacteria. This subfamily, Pol I D, may have a unique role in actinobacterial DNA replication, as I found it to be present together with a standard bacterial PolA, from Pol I C subfamily. Such new subfamilies have recently been identified in PolB enzymes as well (194). Similarities between the different groups may shed light on evolutionary history of these PolA enzymes, whereas conserved motif analysis mapped onto existing structures could hint at differences in polymerase and exonuclease activities.

Being distinguishable from other PolA and similar to PolZ, the DNA polymerase of ϕJL001 suggests that the phage may be subject of DNA modification. Indeed, the genome of ϕJL001 contains T6-like 5hmC glucosyl transferase (protein 22, NCBI ID: AAT69498) and thymidylate synthase (protein 24, NCBI ID: AAT69500). Thus, with high probability it bears a modified (glucosylated) cytosine or thymine,

despite being undetected during phage identification and sequencing (195). Assuming the necessity of an additional hydroxyl group, the modification is expected to be made partially on the level of free nucleotides, similar to T-even phages (147). The ancestor polymerase of ϕVC8 PolZ/ϕJL001 PolA was at least tunable to, or perhaps even suitable for, efficient incorporation of modified nucleotides, pyrimidines and purines alike. Finally, a close relative of ϕJL001, coliphage 9g, contains archaeosine in its DNA that replaces guanine (151, 196). It has, however, a family B DNA polymerase and an AEP primase-polymerase, but no family A DNA polymerase.

3. Phage S-2L Z metabolism: the Z-cluster

In parallel to the previous project and out of curiosity, I initiated and performed the genome re-annotation of the phage S-2L, in order to find there a missing DNA polymerase. I found one that does not belong to the PolA family, but rather to the PrimPol family of AEP superfamily. I proved that it is active as DNA polymerase, however with similar specificity towards dZTP and dATP; I resolved the crystallographic structure of PrimPol's polymerase domain that explains its lack of selectivity. From then on, I did reverse genetics on the phage S-2L, aiming to identify other genes whose products are involved in metabolism of 2-aminoadenine, and in particular its preferential incorporation in the DNA compared to adenine. Functional assays of these proteins of interest were supported by high-resolution crystallographic structures, consisting of the enzymes bound to corresponding nucleotide substrates.

I discovered two genes, *datZ* and *mazZ*, that explain the phenotype of ZTGC-DNA: DatZ is a dATPase (triphosphohydrolase) that removes dATP from the pool of available dNTPs in the host cell during replication, and MazZ hydrolyses GTP to GMP, which is a necessary substrate of PurZ to synthesize a direct precursor of dZMP (Sleiman et al., submitted). Investigation of DatZ and PrimPol constitute the first chapter of the *Results* section, while results on MazZ are described in the second chapter. They are accompanied by a complete functional and structural characterization of S-2L PurZ, unraveling catalysis with alternative nucleotide substrates, including dATP as an energy carrier. In conclusion, genes *datZ*, *mazZ* and *purZ* constitute a tightly-packed Z-cluster involved in 2-aminoadenine metabolism, found in a number of related *Siphoviridae* phages.

It therefore appears that there are at least two (potentially overlapping) strategies to selectively incorporate Z into a phage DNA. One would require a specific DNA polymerase, and potentially a specific helicase, while another modulates the pool of available dNTPs during replication. The latter

one is known to exist in T-even phages since the 1960's (148). Indeed, a conserved datZ gene is found in several other phages with a selective family A DNA polymerase, such as ϕVC8 itself. It remains to be seen if it is the only phosphohydrolase found in nature capable of eliminating dATP from the nucleotide pool.

4. Engineering of PurZ – alternative nucleic acid alphabets

My discovery of the Z-cluster using cyanophage S-2L as a model solves a long-standing mystery of 2-aminoadenine metabolism. Moreover, high resolution crystallographic structures of its members enable potential applications for bioengineering. The active site of PurZ could be potentially redesigned to synthetize novel deoxypurine succinylates, which can be then reduced by a non-specific lyase PurB into final purine analogues. Specifically, a precursor of isoguanine, a guanine with swapped functional groups on positions 2 and 6, could be in principle generated by PurZ from deoxyxanthosine monophosphate (dXMP). This would enable biological synthesis of a base that has already found its application in transcriptable DNA and RNA alphabet extension (171). Ultimately, isoguanine could be used for permanent nucleobase expansion of a living system's DNA, improving upon the synthetic, yet stable X:Y pair (169) by bypassing the need of providing the nucleotide externally. My preliminary HPLC tests on S-2L PurZ with XMP as a substrate indicate that even with the prolonged incubation times the expected product, succinylo-XMP (sXMP), does not appear to be formed. Due to no commercial availability I could not directly test dXMP.

The advantage of PurZ over the related traditional AdSS enzymes involved in IMP to AMP transformation is that it favors dGMP, which is a deoxynucleotide. However, MazZ works readily with both GTP and dGTP, and cellular PurB lyase accepts ribonucleotides as well (197). Through bioengineering, PurZ could in principle be transformed to recognize GMP; with non-specific MazZ, PurB and appropriate kinases, it would open a novel biosynthesis pathway culminating in ZTP. Assuming that this triphosphate can be used for transcription by extant RNA polymerases (or that they can be engineered to do so relatively easily), it would eventually result in ZTGC-RNA. By making the whole pathway inducible, ZTGC-RNA could be generated in a controllable way, enabling its study *in vivo*. Such substitution of the RNA alphabet could have profound consequences on translation at multiple levels, impacting the stability of mRNA, tRNA recognition (for both codons and aminoacyl-tRNA synthetases (198)) and catalytic behavior of rRNA. It could also increase the binding efficiency of siRNA, making for more efficient gene silencers. Finally, ZTGC-RNA could allow for novel RNA folds,

thus creating new ribozymes. The impact of base Z on translation is supported by a recent finding that free 2-aminoadenine extracted from a mushroom *L. inversa* acts as a corrector for a nonsense mutation of two codons (199).

II. Introduction of the S-2L Z-cluster into other species

1. Introduction

The content of the three articles presented in the *Results* section, along with the supplementary data, provides a coherent and rational description of diaminopurine metabolism in a *Siphoviridae* phage clade, to which cyanophage S-2L and vibriophage ϕVC8 belong. It points to the existence of a Z-cluster made of three genes – *datZ*, *purZ* and *mazZ* – that could be sufficient for ensuring A-to-Z exchange in a phage genome. The mechanistic description of their products allows to form a consistent story, similar to the non-standard 5hmC nucleotide pathway described for phages T2, T4 and T6 (147).

Despite the comprehensive molecular investigation of the Z pathway, such an *in vitro* approach has, of course, its limits. Even if the metabolism is accurately represented, it remains to be seen how such a system is regulated *in vivo* inside the virocell. The time-dependence and character of gene transcription (polycistronic or not); the amount of enzyme expression; and its possible regulation through transcription factors or allosteric modulation constitute some of the crucial, yet unanswered points, necessary for detailed description of the biological system. Lastly, related *Siphoviridae* phages with PurZ but different genomic organization (*e.g.* Wayne, Ghobes, Hiyaa) may display a unique 2-aminoadenine processing solution, differing in one or more aspects from the one discovered here in phages S-2L and ϕVC8.

Nevertheless, in spite of this lack of complete information, I also had in mind a famous quote of Richard Feynman, written by him on the blackboard just before his death in 1988:

What I cannot create, I do not understand. (200)

Being convinced that the Z-cluster is a sufficient and standalone unit for base Z synthesis at least in cyanophage S-2L DNA, I tried to employ the molecular knowledge gathered during my thesis' work and transplant the S-2L 2-aminoadenine pathway to other, unrelated organisms. I wanted to alter their DNA composition in order to achieve detectable A-to-Z substitution in their genome. My aim was to learn, through experimentation, if it could be done, or if not, what could be missing or what would need adjusting.

2. Z-cluster in *E. coli*

For the first attempt in Z-cluster transplantation, I have selected *E. coli* BL21-CodonPlus (DE3)-RIPL strain, determined as the system where all S-2L's proteins are readily expressed. Additionally, I had previously found one of the DNA polymerases of *E. coli*, Pol I (Klenow) to incorporate both purines with identical efficiency.

I employed three constructs of crucial S-2L enzymes (Table 8): the first two constructs contained only two genes of the Z-cluster, and the third with the complete set including DatZ dATPase. *E. coli* strains could be selected for the simultaneous presence of two plasmids with different antibiotic resistances (pET100 for ampicillin, pRSF1-Duet for kanamycin).

Construct	#1	#2	#3
DatZ	-	-	pET100
PurZ	pRSF1-Duet (1)	pRSF1-Duet (1)	pRSF1-Duet (1)
MazZ	pET100	pRSF1-Duet (2)	pRSF1-Duet (2)

Table 8. Three constructs with expressible S-2L genes. Numbers in parenthesis stand for the position on the double-slotted plasmid suited for protein co-expression.

I verified the co-expression of enzymes of interest with one-step purification of his-tagged proteins after IPTG induction of the constructs; the results were visualized on SDS-PAGE gels (Fig. 40).

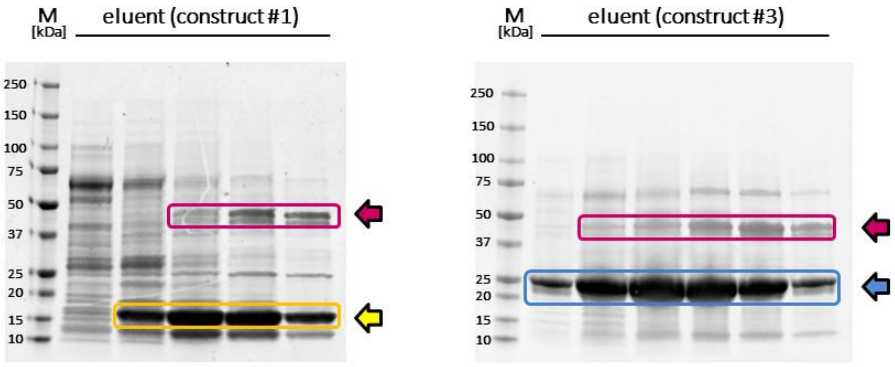


Figure 40. Expression tests on constructs #1 and #3 induced with 840 µM IPTG at OD = 0.6 in 19°C, with products purified on HiTrap Heparin column and visualized on SDS-PAGE gels. Colored arrows stand for S-2L enzymes: DatZ (blue), MazZ (yellow), PurZ (purple). In construct #3 MazZ is invisible, because the second slot of pRSF1-Duet is not his-tagged.

Although S-2L proteins were co-expressed as expected, the bacteria did not seem to survive the process (Fig. 41). Growth curves always came to a halt 1-2 hours after IPTG induction, whether performed in the initial or exponential growth phase; resulting bacterial pellets were white, with sour and slightly irritating smell. Lower IPTG concentration allowed for continued growth, but no S-2L enzymes were detected in such induced cultures.

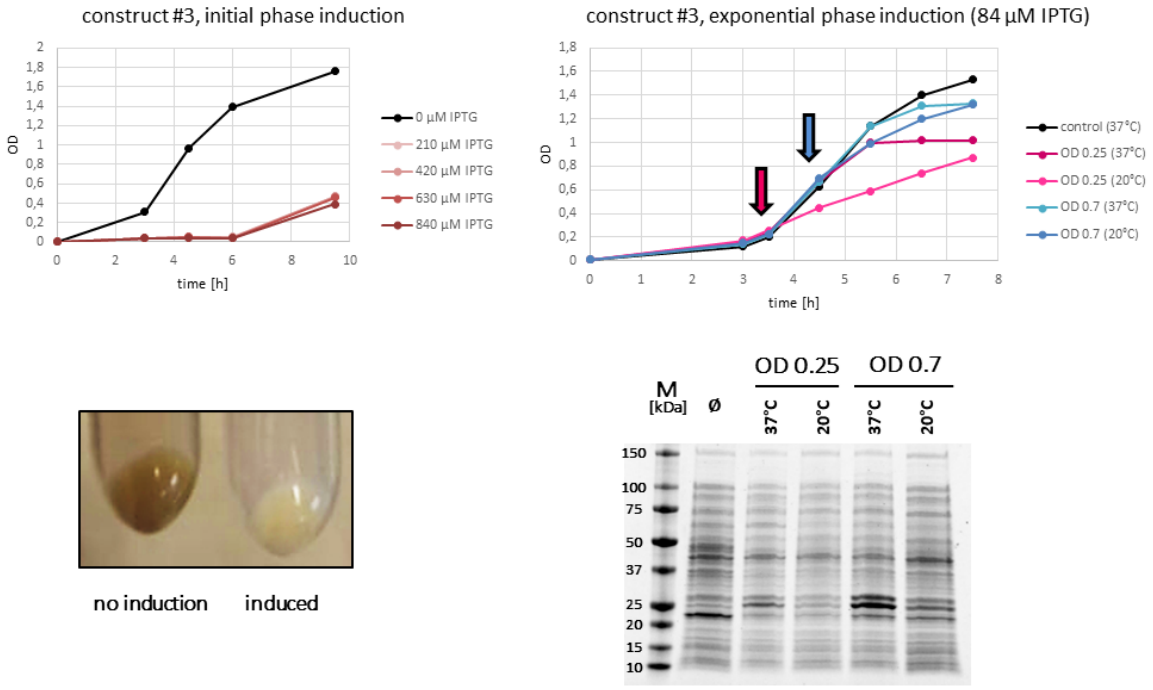


Figure 41. Effects of S-2L enzyme co-expression in *E. coli*. Typical growth curves for initial phase induction (upper left) and phenotype of the resulting bacteria (lower left), similar between all constructs. Lower concentration of IPTG gave better, but still slower growth curves (upper right), but in these trials no S-2L proteins could be identified on SDS-PAGE gels (lower right).

Finally, I tested plasmids of bacteria in arrested growth but with confirmed S-2L protein expression for the presence of 2-aminoadenine. These smaller DNA molecules were chosen because of convenient manipulation techniques and the known role of *E. coli* Pol I in their replication. After isolation, the plasmid sample (construct #3) was digested with benzonase nuclease and analyzed with reverse-phase HPLC using Supelcosil LC-18-S, 5 µm column, with 250 mM NH_4HCO_2 as suspension buffer eluted with a linear gradient of 0-8% acetonitrile at 1 ml min^{-1} flow (201) (Fig. 42). There was no difference between the sample and similarly digested ATGC-DNA control plasmid from non-induced bacteria; additionally, no UV fluorescence was observed in either sample. These preliminary observations suggested that no significant substitution had taken place in the plasmids.

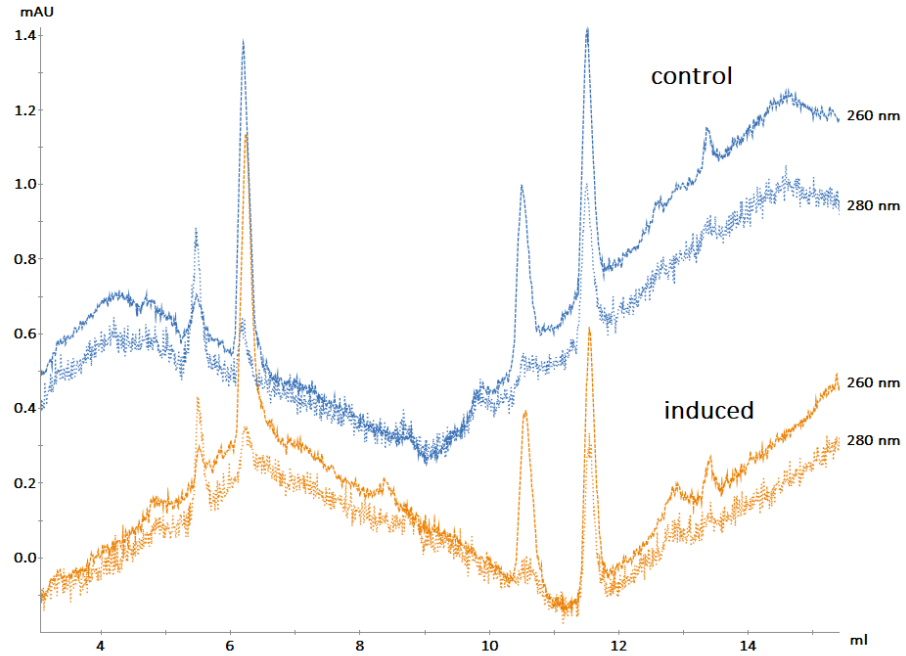


Figure 42. Preliminary HPLC chromatogram of digested plasmids (construct #3), from induced bacteria (orange) and non-induced control (blue). The four major peaks correspond to the individual standard deoxynucleotides (dC, dG, dT, dA). Detection wavelengths are showed on the right.

Although harmless during separate overexpression, S-2L enzymes proved toxic when co-expressed in *E. coli* cells during initial or exponential growth. Similar test for bacteria in stationary phase was not attempted, given that in order to substitute Z for A, their genomic or plasmidic DNA would have to undergo replication. It is possible, owing to the sheer complexity of bacterial DNA transactions, possibly involving nucleobase-specific DNA uncoupling (sensitive to base pair stability) or binding of sequence-specific transcription factors, that even infrequent A-to-Z substitutions can be extremely harmful to a cell. It would mean that bacteria cannot complete DNA replication with such interference.

3. Z-cluster in phage T7

Although unsuccessful, the *E. coli* test provided me with a biochemical setup and knowledge of the system's behavior in bacterial environment. After discussions with another PhD student in virology, Mathieu De Jode, and the subsequent agreement between our supervisors (M. Delarue and L. Debarbieux), we decided to pursue this research in the simpler T7 bacteriophage system.

Taking advantage of the available *E. coli* constructs, we decided to infect the modified bacteria by bacteriophage T7. Being not only a model coliphage with fast replication rate, it also belongs to the same order as S-2L (*Caudovirales*). Additionally, the S-2L genes on the generated plasmids were under the control of T7 promoter; although IPTG induces artificial production of T7 RNA polymerase (ssRNAP), we hoped that the Z-cluster enzymes could be instead expressed during T7 infection.

In the first test, we verified that our T7 strain was indeed attacking *E. coli* BL21-CodonPlus (DE3)-RIPL strain (Fig. 43, left). Next, we went onward to prepare a lysate from bacteria with the S-2L system (construct #3). 500 ml of bacteria grown in LB medium at 37°C, OD = 0.3 were infected with phage T7 (MOI 0.1). After 1h when the lysis was complete, the sample was centrifuged and filtered (0.22 µm), and the lysate supernatant was subjected to one-step histidine-tag purification and SDS-PAGE gel visualization (Fig. 43, right). Unfortunately, despite successful T7 infection, no S-2L proteins were observed, meaning that none of the plasmids were efficiently transcribed.

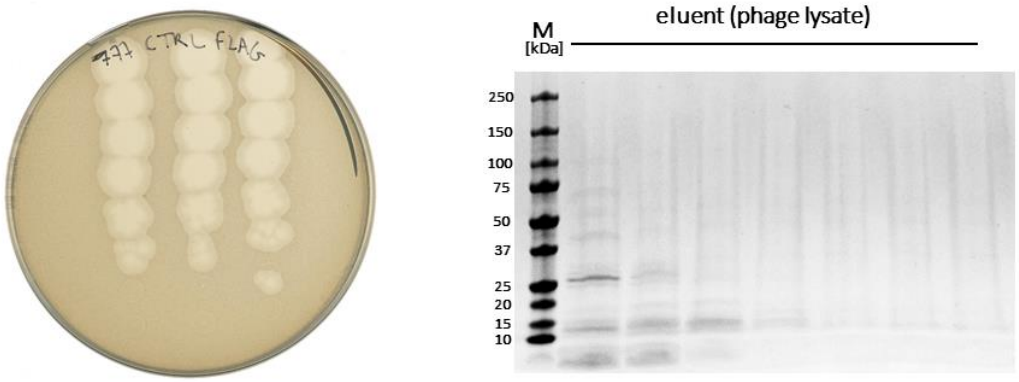


Figure 43. Petri dish covered with *E. coli* BL21-CodonPlus (DE3)-RIPL culture with phage T7 added in decreasing concentrations, top to bottom (left). SDS-PAGE gel of purified T7 lysate using construct #3 (right).

To compensate for this problem, we had to revert to using IPTG. This time, however, we aimed to induce bacteria in the stationary phase: in that setup, if T7 reproduces, it would have to replicate its

DNA in presence of S-2L proteins. Thus, we induced *E. coli* culture (OD = 1.69) in 37°C with 500 µM IPTG and detected overexpressed enzymes after 2h (Fig. 44). Further, we determined in LB agar plate test that the cells were still alive: we observed 41 and 21 colonies 1h and 2h after induction, respectively (dilution 10^4).

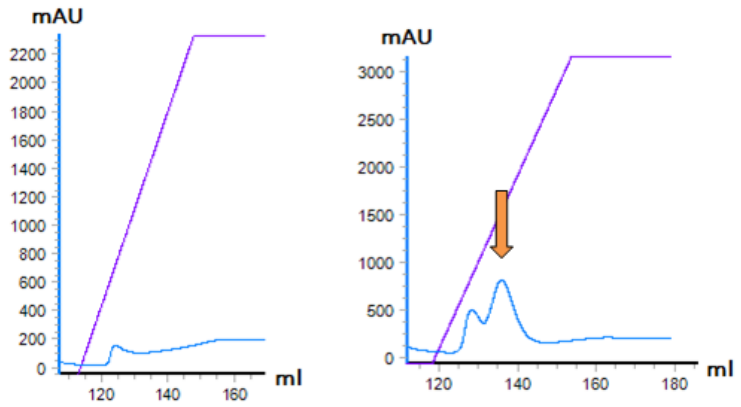


Figure 44. S-2L protein expression in *E. coli*: non-induced (left) and induced during stationary phase (right). The chromatograms were obtained during purification of bacterial lysates on HiTrap Heparin column. The additional peak (orange arrow) corresponds to the his-tagged proteins of interests, DatZ and PurZ.

In a final test, we infected *E. coli* cultures bearing construct #3 in stationary (saturated) phase with phage T7 (MOI 0.2-10), 2h after IPTG induction in 37°C (various concentrations) and followed the dynamics of lysis (Fig. 45). In the small volume used for the automatic 96-well plate reader (180 µl), induction of S-2L proteins did not visibly change the process.

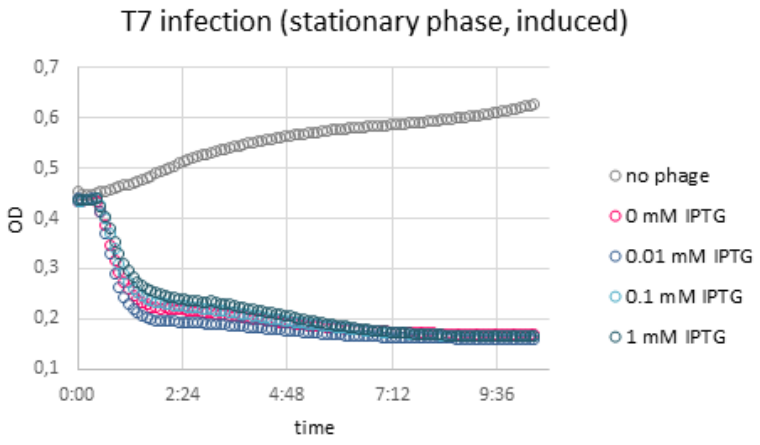


Figure 45. T7 lysis dynamics 2h after induction of *E. coli* cultures (construct #3) in stationary phase (MOI 10).

Ultimately, we successfully collected T7 offspring from *E. coli* cells (construct #3) induced with 1 mM IPTG. We used MOI 1-10 to ensure that only one replication cycle will occur, as we did not know what would be the effect of potential A-to-Z substitution on DNA packing or phage infectiousness in later cycles. After 3h, contrary to the previous experiment and to the control made in parallel, lysis on induced bacteria seemed incomplete (clouded medium and less debris): we attribute the emerging influence of induction on T7 lysis to the bigger volume used (20 ml). Despite this observation, we harvested the phage by centrifuging off the bacterial debris and filtering the supernatant with 0.45 and 0.22 µm filters. Phages from two samples (non-induced and induced) were stored at 4°C.

Next, we extracted the resulting T7 DNA. To avoid *E. coli* nucleic acid contamination, 6 ml of each sample was pre-treated with final concentrations of 120 U of DNase, 240 µg of RNase and 10 mM of $MgCl_2$ at 37°C for 30 min. Reactions were stopped with 20 mM EDTA. Phage capsids and possible protein contaminants were denatured with 100 µg ml^{-1} of proteinase K and 0.5% SDS at 55°C for 30 min. Next, PCI solution in volume ratio 1:1 was added and the sample centrifuged at 1500 g for 5 minutes in room temperature; the step was repeated for the resulting supernatant. 5.6 ml of the aqueous phase was mixed with 620 µl of sodium acetate (3 M, pH 5.2) and 12.4 ml of absolute ethanol. After slow homogenization and 15 min of precipitation on ice, samples were centrifuged at 15000 g for 20 min at 4°C. The supernatant was removed and resulting pellet washed with 3 ml of 70% EtOH. Ethanol was evaporated while avoiding pellet dehydration. The T7 DNA was finally suspended in 250-500 µl of TE buffer yielding a concentration of 117 or 66.4 ng $µl^{-1}$ for control and induced samples, respectively.

As a test for 2-aminoadenine presence in isolated T7 DNA, we have employed a restriction enzyme test (Fig. 46). We used two endonucleases, NcoI and NdeI, known to cut ATGC-DNA sequences, but not ZTGC-DNA (Pezo *et al*., in preparation). Although the digestion pattern corresponds to the one predicted by Serial Cloner 2.6 (202), there was no difference between the induced and control samples, despite short digestion time (15 min). In conclusion, if base Z is present in isolated T7 DNA, only traces of it are actually there (estimated to be less than 1-2%).

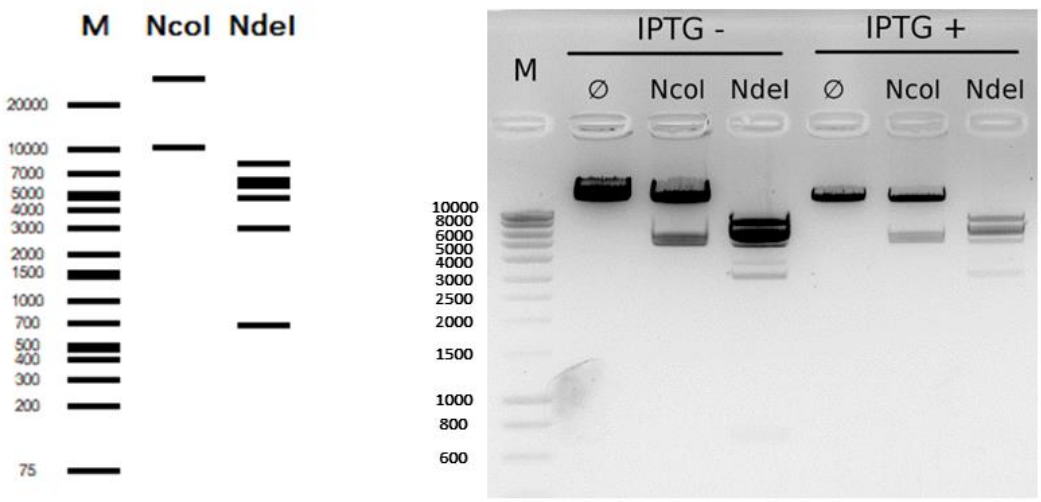


Figure 46. Restriction enzyme assay on T7 DNA; marker lengths are given in basepairs. Virtual prediction with Serial Cloner (left); control and induced samples digested with NcoI and NdeI and visualized on 1% agarose gel (right). First lane in both series on the gel corresponds to non-digested control DNA of T7.

4. Conclusion

Although not successful, preliminary implementation assays of S-2L 2-aminoadenine pathway in *E. coli* or phage T7 constitute a solid foundation onto which further research could be built. The experiments presented here give already a hint on how such a system would behave in a bacterial or orthologous viral environment. It allows us to conclude that the cellular machinery may prove too complex for straightforward introduction of 2-aminoadenine into genomic or plasmidic DNA. Alternatively, there is the possibility that the *E. coli* gene *purB*, whose product completes the reduction of sdGMP into dZMP, should be overexpressed along with Z-cluster enzymes and put under the control of T7 RNA polymerase in a new construct. Because of a lack of time this possibility has not been tested, but all the necessary tools have been nevertheless prepared. Otherwise, because of minimal toxicity effect and optimizable setup, a bacteriophage system constitutes a potential subject for relatively simple synthetic transplantation of base Z into natural DNA.

If successfully introduced, ZTGC-DNA would allow for a thermoresistant, more stable storage of genetic information that can also be faithfully retrieved. Due to a built-in inability to grow if not supplemented with the building block dZTP or its metabolic precursors, synthetic organisms implemented with such genomes could constitute safe tools for biotechnology. They may also prove well-suited for longer survival in conditions at the verge of hospitability, such as outer space, where 2-aminoadenine has indeed been detected (203).

ANNEXES

I. PolA clustering and new subfamilies

1. Introduction

Over the course of my research on a family A DNA polymerase, ϕVC8 PolZ, I was curious how it relates to other known enzymes in terms of phylogeny. Its sequence is close to a DNA polymerase of another described Siphoviridae phage, ϕJL001, which lacks the Z-cluster (195); this relation is additionally confirmed by a signature residue L455 in motif B (ϕVC8 numbering). However, any further phylogenetic connections were obscured by high sequence divergence with PolA representatives, including the ones with resolved crystallographic structure, 9 at the time of writing the present thesis. I realized that PolZ/PolA of ϕVC8 and ϕJL001 are unique and correspond to a novel clade of family A DNA polymerases found in marine viriplankton (204), that stay however unrelated to any subfamily described in the literature (205, 206) or found at NCBI's Conserved Domain Database (CDD) (207) (Fig. 47).

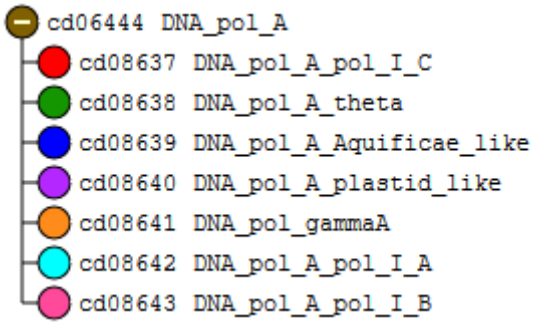


Figure 47. Distribution of DNA polymerases from family A (PolA) into subfamilies, from NCBI CDD as of the time of writing the present thesis (207). They include bacterial Pol I subfamilies (A, B, C), eukaryotic DNA Pol γ, θ and ν, plant plastidic DNA polymerase and Aquificae-like PolA (found in bacteria, viruses and Apicomplexa). PolZ/PolA of ϕVC8 and ϕJL001 stay distinct from these major groups.

Thus, my secondary aim was to redistribute all known family A DNA polymerases into subfamilies, in hope of finding new ones and determining where ϕVC8 PolZ/ϕJL001 PolA are placed among them. To that end, I identified all sequences in Uniprot database (208) labelled as PolA (PFAM ID: PD00476). However, such richness of data – 33558 seuqences as of 27 June 2018 – was often redundant or incomplete. To sort the entries, I wrote a program in Python, eliminating sequences of length less than 500 amino acids, incomplete (with character "X") or redundant, removing sequences with

identical 4 first amino acids in a crude approach. This reduced the dataset to 3640 unique PolA representatives.

This number was still too big to make a correct multialignment and construct an informative phylogenetic tree. It is due not only to the data size and computational time demand, but also high divergence of the dataset, lowering the accuracy of most available multialignment methods, such as progressive alignment implemented in ClustalW (209). However, as an alternative, a clustering approach based on pairwise similarity called CLANS (210) was recently implemented to re-assess the diversity of AEP superfamily and family B DNA polymerases (67, 194) (Fig. 48). This method simulates a general repulsive force between the sequences visualized in 3D space by a set of points, and translates pairwise similarity to an attractive force between them. Such simulation results in similar sequences being clustered together, but away from other, more distant ones. Thus, CLANS analysis is well-suited for a large number of divergent sequences, allowing for their clear separation.

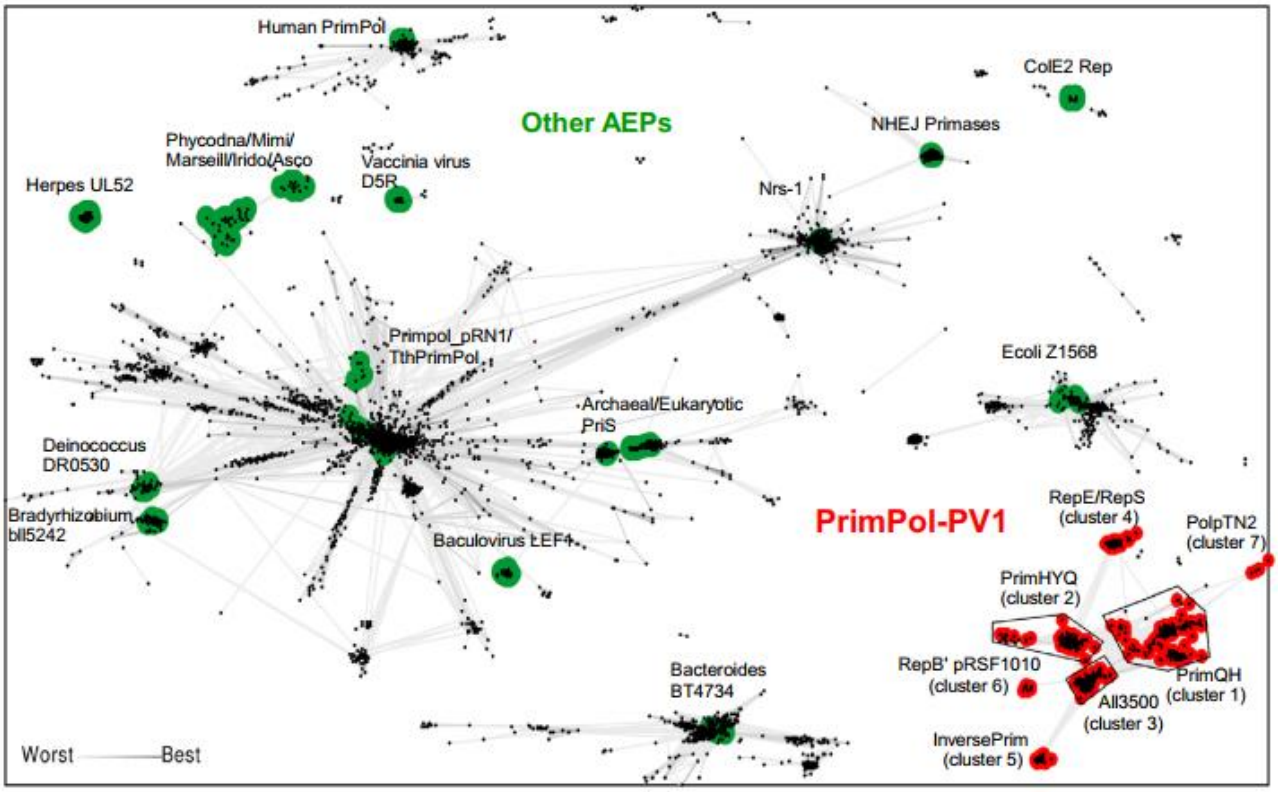


Figure 48. Clustering of AEP superfamily, including PrimPol family (adapted from (8)). Each point is a unique AEP sequence, that may cluster with others in subfamilies (green and red).

2. Results

I ran a CLANS simulation with default parameters (200 rounds) on the dataset of 3640 representative PolA sequences. Resulting 3D sequence distribution was then automatically clusterized with an accompanied tool (Network-based clustering, minimum 40 sequences per cluster) and rendered with a color code (Fig. 49).

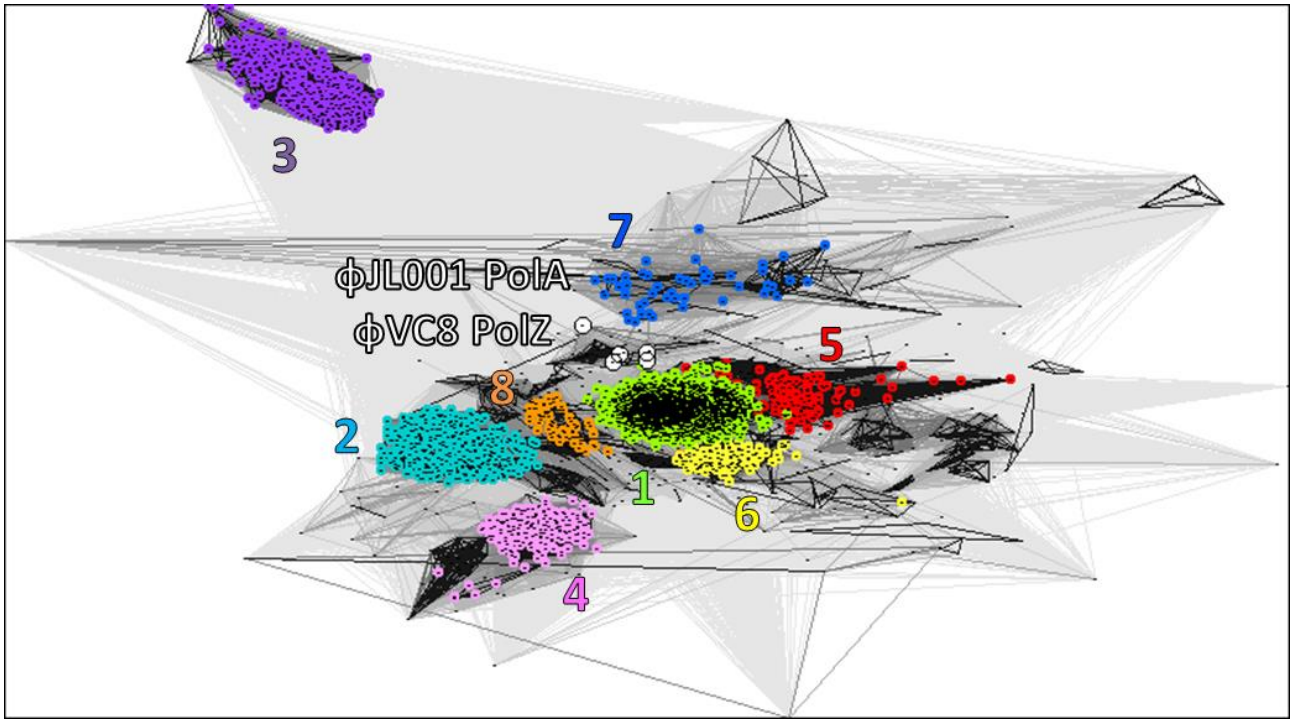


Figure 49. 2D projection of a 3D distribution of representative family A DNA polymerase sequences, divided into 8 major clusters (colored dots). Link saturation correlate with strength of phylogenetic connection between the sequences. Representative sequences of close ϕVC8 PolZ homologues and ϕJL001 PolA are marked by white circles, outside of 8 identified clusters.

Almost all PolA clusters identified by CLANS correspond to the ones found in NCBI CDD. Although not part of an identified cluster, Human DNA polymerase ν (PDB ID: 4XVK) is found close to cluster theta, which is expected from their relatively close homology (211). Additionally, Plasmodium falciparum apicoplast PolA (PDB ID: 5DKU) is situated next to cluster 8 (Aquificae-like).

Importantly, cluster 2, the second biggest one with 246 representative sequences, does not correspond to any of the previously described subfamilies despite its clear separation in clustering analysis. It contains PolA sequences of consistent length, of average 564 (±27) aa, with polymerase and 3'-5' exonuclease domains. All 77 species identified there belong to various orders in phylum

Actinobacteria, with the predominant ones being Streptomyces (46 subspecies), Rhodococcus (18 ss), Arthrobacter (17 ss) and Microbacterium (16 ss). In a preliminary search, I have found that Streptomyces sp. CT34 genome contains a traditional PolA from major bacterial Pol I C subfamily (NCBI reference WP_043263881), but also another PolA belonging to cluster 2 (WP_043265455). Thus, the latter does not seem to replace the common Pol I version and may be linked to different cellular functions than Okazaki fragment maturation (212) and plasmid replication (93).

Characteristics of each cluster, including corresponding CDD subfamily (Fig. 47) and available representatives with resolved crystallographic structure, are gathered in Table 9.

Cluster	Number of sequences	Corresponding subfamily (CDD)	Known structures (PDB ID)
1 (lime)	2037	Pol I C (cd08637)	*E. coli* (1D8Y), *G. stearothermophilus* (1L3S), *T. aquaticus* (1TAQ), *M. smegmatis* (6VDD)
2 (cyan)	246	-	-
3 (purple)	232	gamma (cd08641)	*H. sapiens* γ (3IKM)
4 (pink)	152	Pol I A (cd08642)	-
5 (red)	98	theta (cd08638)	*H. sapiens* θ (4X0Q), *H. sapiens* ν (adjacent) (4XVK)
6 (yellow)	84	plastid-like (cd08640)	-
7 (blue)	49	Pol I B (cd08643)	phage T7 (1SKR)
8 (orange)	48	*Aquificae*-like (cd08639)	*P. falciparum* (adjacent) (5DKU)

Table 9. PolA subfamily distribution. Cluster 2 of actinobacterial PolA does not correspond to any known subfamily.

3. Discussion

Indeed, I found that ϕVC8 PolZ/ϕJL001 PolA are not part of any major subfamily; likewise, there are too few representative sequences at the present time in this putative new subfamily to be automatically recognized as a cluster in itself. However, they stay connected and visibly separate, confirming their close and distinct phylogenetic relationship (Fig. 49). They place themselves next to the biggest group of bacterial PolA (CDD Pol I C), which most probably translates to their phylogenetic closeness as well. A logical scenario would be that the ancestor of PolZ diverged most probably from bacterial PolA (Pol I C subfamily), that further speciated from a common ancestor with ϕJL001 PolA, acquiring its dZTP specificity along the way or later. Functional tests on PolA from the clade of ϕJL001 could help link common signature features to catalytic activities observed for ϕVC8 PolZ.

Moreover, the newly identified PolA cluster 2, a new actinobacterial PolA subfamily (which I call Pol I D) could be of special interest for structural and functional enzymatic studies due to its previously undescribed separation from other classes. Two distinct variants of PolA existing simultaneously in Streptomyces sp. CT34 evoke two forms of family C DNA polymerase, PolC and DnaE of different (leading vs lagging) strand preference during chromosome replication (87, 213). Both their genes were identified in Actinobacteria, but also other phyla of Gram-positive bacteria. As the Okazaki fragments are found exclusively on the lagging strand, the role of Pol I D subfamily of PolA may prove to be situational or completely unique.

II. Fast and efficient purification of SARS-CoV-2 RNA dependent RNA polymerase complex expressed in *Escherichia coli*

Madru C, Rosario S, Tekpinar A, Czernecki D, Brûlé S, Sauguet L and Delarue M

submitted to *PLoS One* (August 2020) – under revision

1. Preface

As stated previously, I have contributed to the work where me and my colleagues describe an efficient expression system of an active RNA-dependent RNA polymerase (RdRP) of SARS-CoV-2. I provided the technical support for polymerization activity assays in ribonuclease-free (RNase-free) conditions.

Beyond the scope of the article, I have also participated in commercial tests of RdRP potential inhibitors, taking advantage of the technical setup and previous expertise.

Fast and efficient purification of SARS-CoV-2 RNA dependent RNA polymerase complex expressed in Escherichia coli

Clément Madru[1], Ayten Tekpinar[1], Sandrine Rosario[1], Dariusz Czernecki[1,2], Sébastien Brûlé[3], Ludovic Sauguet[1]* and Marc Delarue[1]*

[1]Unit of Structural Dynamics of Macromolecules, Institut Pasteur & CNRS UMR 3528, Paris (France).

[2]Sorbonne Université, École Doctorale Complexité du Vivant (ED515), Paris (France).

[3]Molecular Biophysics Platform, C2RT, Institut Pasteur, CNRS UMR 3528, Paris (France).

[*]ludovic.sauguet@pasteur.fr (LS); *marc.delarue@pasteur.fr (MD)

ABSTRACT

To stop the COVID-19 pandemic due to the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which caused more a million deaths to date, new antiviral molecules are urgently needed. The replication of SARS-CoV-2 requires the RNA-dependent RNA polymerase (RdRp), making RdRp a very good target for antiviral agents. RdRp is a multi-subunit complex composed of 3 viral proteins named nsp7, nsp8 and nsp12, that ensure the transcription and replication of the ~30 kb RNA genome. The main strategies employed so far for the overproduction of RdRp consists in expressing and purifying the 3 subunits separately before assembling the complex *in vitro*. Moreover, nsp12 shows limited solubility in bacterial expression systems and is often produced in insect cells. Here, we describe an alternative strategy to produce the SARS-CoV-2 RdRp in *E. coli*, using a single plasmid. Characterization of the purified recombinant SARS-Cov-2 RdRp shows that it forms a tetramer with the expected stoichiometry. RNA polymerization activity was measured using primer-extensions assays showing that the purified enzyme is functional. The purification protocol can be achieved in one single day. Our construction has been made available to the entire scientific community through the addgene plasmid depository.

INTRODUCTION

The COVID-19 pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has affected millions of people with a death toll amounting to hundreds of thousands worldwide (1–4). SARS-CoV-2 is an enveloped, single-stranded virus with a positive-sense RNA genome (5,6). One of the most promising druggable targets is the RNA-dependent RNA polymerase (RdRP), which is a central element of SARS-CoV-2 life cycle, responsible for the transcription and replication of the ~30kb genome (7–10). RdRp are targeted in different RNA viruses by an important group of antiviral agents including nucleotides and nucleosides analogs (11–13). So far, Remdesivir, Favipiravir, Ribavirin, Galidesivir, and EIDD-2801, have been shown to efficiently inhibit SARS-CoV-2 replication in cell-based assays (14–17) but their efficiency in humans remains to be assessed, rendering the search for new inhibitors still of interest. RdRp is composed of 3 viral non-structural proteins (nsp) named nsp7, nsp8 and nsp12. The core component nsp12 hosts the catalytic polymerase activity (18) which is greatly enhanced by the two accessory subunits nsp7 and nsp8 (8,19). Recently, RdRp has been a subject of intensive structural biology efforts, yielding high resolutions cryo-EM structures of the RdRp apo form(20,21), bound with RNA (22,23), and with inhibitors (23,24).

The production of sufficient amounts of heterologous RdRp with a native structure and full biological activity is a prerequisite for the discovery, optimization and comprehensive evaluation of new drugs directed against SARS-CoV-2, including in High Throughput Screening (HTS) assays involving very large chemical libraries (25). The main strategies employed so far for the overexpression of recombinant RdRp consists in expressing and purifying the 3 subunits separately before assembling the complex *in vitro* (19,22). Moreover, while nsp7 and nsp8 express readily in *Escherichia coli*, nsp12 shows limited solubility in bacterial expression systems and is often produced in insect cells (20,22,24). These approaches

lengthen the protein expression and purification steps, making RdRp isolation cost- and time-consuming.

Here, we describe an alternative strategy to produce the SARS-CoV-2 RdRp directly in *E. coli*, using a single plasmid. Characterization of the purified recombinant SARS-CoV-2 RdRp using Matrix-Assisted Laser Desorption/Ionization (MALDI-TOF/TOF), negative-stain electron microscopy, and analytical ultracentrifugation assays (AUC) revealed that it forms a tetramer with the expected stoichiometry. RNA polymerization activity was measured using primer-extensions assays and showed that the purified enzyme is indeed functional. This approach provides a useful alternative to more expensive and complicated protein expression systems, and offers many practical advantages inherent to bacterial systems, such as easy generation of mutants and simple cultivation handling. Our fast, single-day purification protocol resulted in a stable and active complex that can be routinely used in most protein biochemistry laboratories for drug screening, as well as for functional and structural studies.

MATERIALS AND METHODS

Cloning

The open reading frames (ORF) of the nsp7, nsp8 and nsp12 genes from SARS-CoV-2 virus were synthesized commercially by geneArt (Thermo Fisher) and inserted into a modified pRSFDuet-1 (Novagen), containing kanamycin-resistance and lacI-encoding genes **(Fig1A)**. The vector encodes two multiple cloning sites (MCS1 and 2), each of which is preceded by a T7 promoter, lac operator, and ribosome binding site (RBS). Nsp12 was inserted in the MCS1 between BamH1 and Not1, preceded by the Tobacco Etch Virus (TEV) protease-recognition sequence ENLYFQG fused to a 14xhistidine tag (26). The DNA coding for nsp7 and nsp8 was cloned as a unique cassette in the MSC2 between NdeI and XhoI, and an RBS sequence has

been inserted between the two genes **(Fig1B)**. Plasmid sequence was confirmed by DNA sequencing (Eurogentec).

Transformation of competent cells

100 ng of pRSFDuet-1(nsp12)(nsp7)(nsp8) were added to 50 µL chemically competent BL21-CodonPlus(DE3)-RIPL strain from *E. coli* (Agilent Technologies) and incubated for 20 min on ice. Cells were heat-shocked 30 s at 42°C and incubated on ice for 2 min. 900 µL of super optimal broth (SOC) medium were added and the mixture incubated at 37°C for 1 h. 200 mL of yeast extract and tryptone (2YT) medium supplemented with 100 µg/mL kanamycin and 50 µg/mL chloramphenicol were then inoculated with the transformation reaction in 1 L flask, and cells were grown at 37°C overnight with 180 rpm agitation.

Protein expression

30 mL of the overnight culture were inoculated into 1.2 L of yeast extract and tryptone (2YT) medium supplemented with 100 µg/mL kanamycin and 50 µg/mL chloramphenicol in 5 L flasks and incubated at 37°C with 180 rpm agitation for a few hours until OD_{600} reaches 0.6. Flasks were then placed for 15 min at 4°C and recombinant protein expression was induced by adding 0.1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) and incubating over-night at 18°C with 180 rpm agitation. Cells were harvested by centrifugation, washed once with fresh 2YT, and stored at -80˚C.

Protein purification

Cells were resuspended in the HisTrap buffer A (50 mM Na-HEPES at pH 8, 500 mM NaCl, 10 mM imidazole) supplemented with complete EDTA-free protease inhibitors (Thermo Fisher) and 500 units of benzonase (Sigma) at 4˚C. Resuspended cells were then lysed by

mechanical disruption with 3 passes through a pre-cooled cell disruptor (Constant System Limited) at 1.4 kPa, and the lysate was centrifuged at 20 000 g for 30 min at 4˚C. All the following steps described below were performed with chromatography columns from GE Healthcare connected to an ÄKTA pure system (GE Healthcare) at 4˚C. After centrifugation, the clear supernatant containing the RdRp complex was loaded onto a 5-mL HisTrap. The column was then washed with 25mL of 5% Histrap buffer B (50 mM Na-HEPES pH 8, 500 mM NaCl, 500 mM imidazole). The complex was finally eluted using a 50 mL linear gradient of imidazole (5%-100% Histrap buffer B). Fractions were analyzed by SDS-PAGE 4-20%, and DNA contamination was detected by measuring the ultraviolet (UV) absorption spectra and the ratio of absorbance at 260 nm vs 280 nm (A260/280). RdRp-containing HisTrap fractions were combined and 5-fold diluted in a 50mM Na-HEPES pH 8 solution, before being loaded onto a 5-ml heparin HiTrap HP pre-equilibrated in the Heparine buffer A (50 mM Na-HEPES pH 8, 150 mM NaCl). The column was washed with 25mL of Heparine buffer A, and the protein complex was eluted with a 50mL linear gradient of NaCl realized by mixing Heparin buffer A with Heparin buffer B (50 mM Na-HEPES pH 8, 600 mM NaCl). Proteins fraction purity was analyzed by SDS-PAGE 4-20%, and by UV spectra measurement showing an A260/A280 ratio of 0.6. The purest fractions containing RdRp complex were combined and concentrated up-to 5 mg/mL using Amicon Ultra-4 centrifugal filter units 30 000 NMWL (EMD Millipore). The purification was completed using a size-exclusion chromatography on a Superdex 200 10/300 equilibrated in S200 buffer (20 mM Na-Hepes pH 8, 300 mM NaCl, 1 mM MgCl2). RdRp containing fractions were then combined, concentrated up-to 3 mg/mL, flash-frozen in liquid nitrogen, and stored at -80 ˚C.

Analytical ultracentrifugation assays

Sedimentation velocity experiment was performed with a Beckman Coulter Optima analytical ultracentrifuge (Beckman-Coulter, USA) with an An-60 Ti rotor at 20°C. The RdRp complex at a concentration of 1.4 mg/ml was centrifuged at 42,000 rpm in 3-mm double-sector epoxy centerpieces. 100 scans were collected at 1-min intervals with a radial step size of 0.001cm. Detection of the protein complex as a function of radial position and time was performed by absorbance measurements at 250 nm, 280 nm and by interference detection. Profiles were analyzed using the continuous (s) distribution model of the software Sedfit (27). The partial specific volume of the protein, the viscosity and the density of the buffer at 20°C were theoretically calculated with the software Sednterp (Spin Analytical, USA).

Negative-stain electron microscopy

RdRp at 0.05 mg/mL in buffer E was deposited on glow-discharged carbon-coated copper grids CF400-CU (Electron Microscopy Sciences) and contrasted 1 minute in 2% uranyl acetate. Data collection was performed using a Tecnai biotwin T12 (Thermo Fisher) equipped with a LaB6 filament, operating at 120 keV. Images were recorded using an Eagle camera (Thermo Fisher) at a nominal magnification of 49000, using a 3µm defocus.

RNA Primer extension assays

The RNA duplex was prepared by mixing a 40-mer RNA template (5'-CUAUCCCCAUGUGAUUUUAAUAGCUUCUUAGGAGAAUGAC-3') corresponding to the 3' end of SARS-CoV-2 genome with a 20-mer fluorescent RNA primer (5'-FAM-GUCAUUCUCCUAAGAAGCUA-3') in water. The mix was then heated 2 min at 70 °C and slowly cooled to room temperature. The primer extension assay was performed with 500 nM of purified RdRp complex in the presence of 500 µM NTPs and 100 nM RNA duplex, in a reaction buffer containing 20 mM Na-Hepes pH 8, 50 mM NaCl, 3 mM and MgCl2. Reaction

was carried out in 100 µL for 60 minutes at 30 °C. After 5, 10, 30 and 60 min, 15 µL were pipetted from the reaction mix and immediately diluted in 2X TBE-urea sample buffer containing Xylene Cyanol and bromophenol blue. Reaction was then stopped by boiling samples for 3 min at 90°C. Products were separated on a TBE-urea gels 15% (Novex). Images were taken using a typhoon FLA 9000 (GE Healthcare).

RESULTS AND DISCUSSION

Cloning and expression of the SARS-CoV-2 RdRp

A single plasmid was employed to co-express the nsp7, nsp8 and nsp12 subunits of the SARS-CoV-2 RdRp in *E. coli* **(Fig 1)**. The genes encoding for nsp7 and nsp8 were optimized for bacterial expression. However, the NSP12 gene was initially designed and synthetized for production in insect cells. Accordingly, the BL21-CodonPlus(DE3)-RIPL strain, containing extra copies of rare tRNA genes , was used.

The protein induction conditions were optimized by testing various IPTG concentration culture media type, incubation duration and temperature **(FigS1)**. Results showed that media composition had a remarkable effect on the expression of recombinant RdRp, whereas IPTG concentration, duration and temperature did not have a significant influence on the recombinant expression level. However, the IPTG concentration was reduced to the lowest value because had a negative effect on cell growth. The highest ratio of recombinant RdRp in cell lysate per liter of culture was obtained with 0.1 mM of IPTG at 20°C in 2YT media after overnight incubation.

Purification of the SARS-CoV-2 RdRp

The purification protocol consists of three successive chromatographic steps, including nickel affinity, heparin and size exclusion columns. The catalytic subunit nsp12 is fused to an N-terminal 14His-tag, enabling large-scale purification of the recombinant RdRp using nickel affinity chromatography **(Fig 2A)**. The UV spectra of the eluted fractions showed a large DNA contamination with A260/A280 ratio of 1.4 at this stage. The following heparin affinity chromatography was sufficient to remove this DNA contamination, reducing this ratio to 0.6 **(Fig 2B)**. Finally, the purification was completed using a size-exclusion chromatography which showed one single peak, reflecting the homogeneity of the sample **(Fig 2D)**. The SDS-PAGE analysis showed 3 bands at 110 kDa, 22 kDa and 9 kDa, corresponding respectively to 14His-nsp12, nsp8 and nsp7, as determined through in-gel tryptic digestion coupled with Matrix-Assisted Laser Desorption/Ionization (MALDI-TOF/TOF) analysis. An additional band corresponding to an approximately 12 kDa protein was identified as the globular C-terminal part of nsp8 (nsp8-CTD), showing that the N-terminal region of nsp8 is partially proteolyzed (~80 amino-acids). Recent structural studies have revealed that this N-terminal region of nsp8 is flexible and gets ordered when the RNA duplex exits from the enzyme's active site (22). Yet, this region is not required for RNA polymerase activity but improves its processivity by perpetuating the interactions with the RNA backbone (22). An overall yield of 1.1 mg RdRp complex was obtained per litter of *E. coli* pellet after the final size exclusion chromatography. Depending on the applications, the N-terminal 14His-tag fused to nsp12 may be removed following TEV-protease cleavage.

Biophysical and functional characterization of the purified RdRp complex

Analytical ultracentrifugation revealed that the purified RdRp is homogeneous, exhibiting a main peak, with a sedimentation coefficient of 6.2 S. The frictional ratio (f/f0) of 1.5 suggests that the complex has an extended shape, consistent with the RdRp structure **(Fig**

1C) (22,23). The combination of the sedimentation coefficient and the peak signal at 250 nm, 280 nm as well as the interference signal allows to resolve the complex stoichiometry using our in-house routine. The results are compatible with the expected (nsp12)-(nsp7)-(nsp8)-(nsp8-CTD) stoichiometry (**Fig 3A**). In addition, the purified RdRp complex was applied onto glow-discharged carbon coated EM grids, and stained with a 2% uranyl formate aqueous solution. Images were recorded with a defocus of -3µm, at the instrument magnification of 49000. The negative-staining EM images show discrete particles of uniform size indicating that the purified sample is homogeneous and not aggregated **(Fig 3B).** Finally, in order to verify whether the purified RdRp is active, an RNA primer extension assay was performed. The purified RdRp was incubated for 1 h with a 40-mer RNA template mimicking the 3'-extremity of the viral genome, primed by a fluorescently-labelled 20-mer RNA. As shown in **Fig 3C**, the purified RdRp is active, and mediates primer-dependent RNA elongation. Moreover, both fresh and thawed protein sample showed similar activity, suggesting that freezing does not affect enzyme activity (data not shown). In summary, the quality of the purified RdRp has been fully assessed using biophysical and biochemical assays. The sample has the expected stoichiometry, is homogeneous, and is functionally active.

CONCLUSION

Motivated by *E. coli*'s broad accessibility, ease of culture, rapid growth rates, and proven scalability, we developed an efficient expression and purification system for the SARS-CoV-2 RdRp complex in this bacterial host. Our construction has been made available to the entire scientific community through the Addgene plasmid repository (Addgene ID: 159133). The purification protocol can be achieved in one single day. The resulting sample is pure, homogeneous, properly active, and therefore suitable for drug screening, extensive site-directed mutagenesis, as well as for functional and structural studies.

ACKNOWLEDGMENTS

The negative-staining microscopy data were collected at the Ultrastructural BioImaging Facility (UtechS UBI – Institut Pasteur). We would like to thank Gérard Pehau-Arnaudet (UtechS UBI – Institut Pasteur). for help with sample preparation and data collection, and Dr. Bertrand Raynal (Molecular Biophysics Platform, C2RT, Institut Pasteur) for helping us with AUC data analysis. We wish also to thank Dr. Margarida Gomes for helpful advices for molecular biology. This work was funded by the Institut Pasteur, and by an ANR JCJC grant ANR-17-CE11-0005-01. The post-doctoral fellowship of C.M is funded by the Pasteur-Roux-Cantarini fellowship from the Institut Pasteur. The fellowship of D.C was funded by Sorbonne University ED515.

AUTHOR CONTRIBUTIONS

Conceptualization: Clément Madru, Ludovic Sauguet, Marc Delarue.

Funding acquisition: Ludovic Sauguet, Marc Delarue

Investigation: Clément Madru , Sandrine Rosario, Dariusz Czernecki, Sébastien Brûlé.

Methodology: Ludovic Sauguet.

Project administration: Clément Madru.

Supervision: Marc Delarue.

Validation: Ludovic Sauguet, Marc Delarue.

Writing – Original Draft Preparation: Clément Madru.

Writing – Review & Editing: Clément Madru, Sébastien Brûlé, Ludovic Sauguet, Marc Delarue.

REFERENCES

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020 Feb;395(10223):497–506.

2. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar;579(7798):270–3.

3. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265–9.

4. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May;20(5):533–4.

5. Hilgenfeld R, Peiris M. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. Antiviral Res. 2013 Oct;100(1):286–95.

6. Neuman BW, Buchmeier MJ. Supramolecular Architecture of the Coronavirus Particle. In: Advances in Virus Research [Internet]. Elsevier; 2016 [cited 2020 Jul 29]. p. 1–27. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065352716300446

7. Snijder EJ, Decroly E, Ziebuhr J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. In: Advances in Virus Research [Internet]. Elsevier; 2016 [cited 2020 Jul 28]. p. 59–126. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065352716300471

8. Posthuma CC, te Velthuis AJW, Snijder EJ. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. Virus Res. 2017 Apr;234:58–73.

9. Sheahan TP, Sims AC, Graham RL, Menachery VD, Gralinski LE, Case JB, et al. Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses. Sci Transl Med. 2017 Jun 28;9(396):eaal3653.

10. Buonaguro L, Tagliamonte M, Tornesello ML, Buonaguro FM. SARS-CoV-2 RNA polymerase as target for antiviral therapy. J Transl Med [Internet]. 2020 Dec [cited 2020 Aug 3];18(1). Available from: https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-020-02355-3

11. Subissi L, Imbert I, Ferron F, Collet A, Coutard B, Decroly E, et al. SARS-CoV ORF1b-encoded nonstructural proteins 12–16: Replicative enzymes as antiviral targets. Antiviral Res. 2014 Jan;101:122–30.

12. Jordan PC, Stevens SK, Deval J. Nucleosides for the treatment of respiratory RNA virus infections. Antivir Chem Chemother. 2018 Jan;26:204020661876448.

13. Elfiky AA. Anti-HCV, nucleotide inhibitors, repurposing against COVID-19. Life Sci. 2020 May;248:117477.

14. Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res. 2020 Mar;30(3):269–71.

15. Cai Q, Yang M, Liu D, Chen J, Shu D, Xia J, et al. Experimental Treatment with Favipiravir for COVID-19: An Open-Label Control Study. Engineering [Internet]. 2020 Mar [cited 2020 Jul 28]; Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095809920300631

16. Lu C-C, Chen M-Y, Lee W-S, Chang Y-L. Potential therapeutic agents against COVID-19: What we know so far. J Chin Med Assoc. 2020 Jun;83(6):534–6.

17. Sheahan TP, Sims AC, Zhou S, Graham RL, Pruijssers AJ, Agostini ML, et al. An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. Sci Transl Med. 2020 Apr 29;12(541):eabb5883.

18. Ahn D-G, Choi J-K, Taylor DR, Oh J-W. Biochemical characterization of a recombinant SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA templates. Arch Virol. 2012 Nov;157(11):2095–104.

19. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, et al. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proc Natl Acad Sci. 2014 Sep 16;111(37):E3900–9.

20. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 2020 May 15;368(6492):779–82.

21. Peng Q, Peng R, Yuan B, Zhao J, Wang M, Wang X, et al. Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2. Cell Rep. 2020 Jun;31(11):107774.

22. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. Structure of replicating SARS-CoV-2 polymerase. Nature [Internet]. 2020 May 21 [cited 2020 Jul 28]; Available from: http://www.nature.com/articles/s41586-020-2368-8

23. Wang Q, Wu J, Wang H, Gao Y, Liu Q, Mu A, et al. Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. Cell. 2020 Jul;182(2):417-428.e13.

24. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science. 2020 Jun 26;368(6498):1499–504.

25. Sugiki T, Fujiwara T, Kojima C. Latest approaches for efficient protein production in drug discovery. Expert Opin Drug Discov. 2014 Oct;9(10):1189–204.

26. Raia P, Carroni M, Henry E, Pehau-Arnaudet G, Brûlé S, Béguin P, et al. Structure of the DP1–DP2 PolD complex bound with DNA and its implications for the evolutionary history of DNA and RNA polymerases. Stock AM, editor. PLOS Biol. 2019 Jan 18;17(1):e3000122.

27. Schuck P, Perugini MA, Gonzales NR, Howlett GJ, Schubert D. Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems. Biophys J. 2002 Feb;82(2):1096–111.

FIGURES

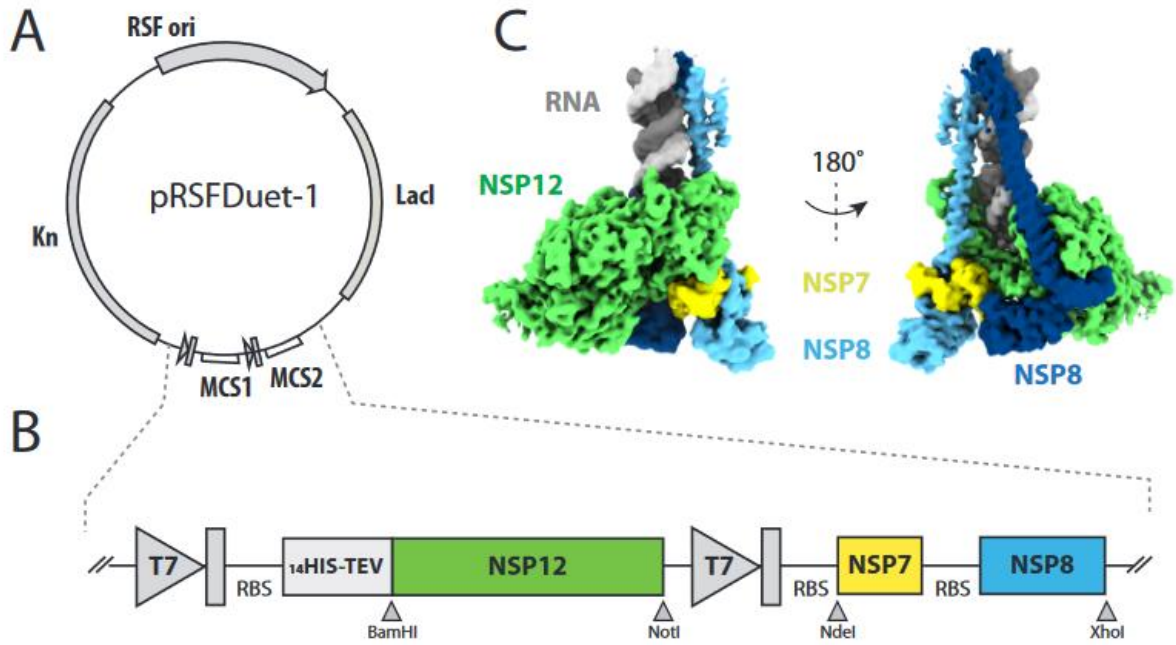


Figure 1: Recombinant expression vector of the SARS-CoV-2 RdRp complex

(A) Plasmid map of the pRSFDuet-1 used in this study. The vector encodes two multiple cloning sites (MCS) each of which is preceded by a T7 promoter, a lac operator and a ribosome binding site (RBS). **(B)** Focus on MCS1 and MCS2. Nsp12 was inserted in the MCS1 between BamH1 and Not1, preceded by a TEV-cleavable 14xhistidine tag. The DNA coding for nsp7 and nsp8 was cloned as a unique cassette in the MSC2 between NdeI and XhoI, and an RBS sequence has been inserted between the two genes. **(C)** Two orthogonal views of the cryo-EM structure of the SARS-CoV-2 RdRp (PDB code: 6YYT) (22).

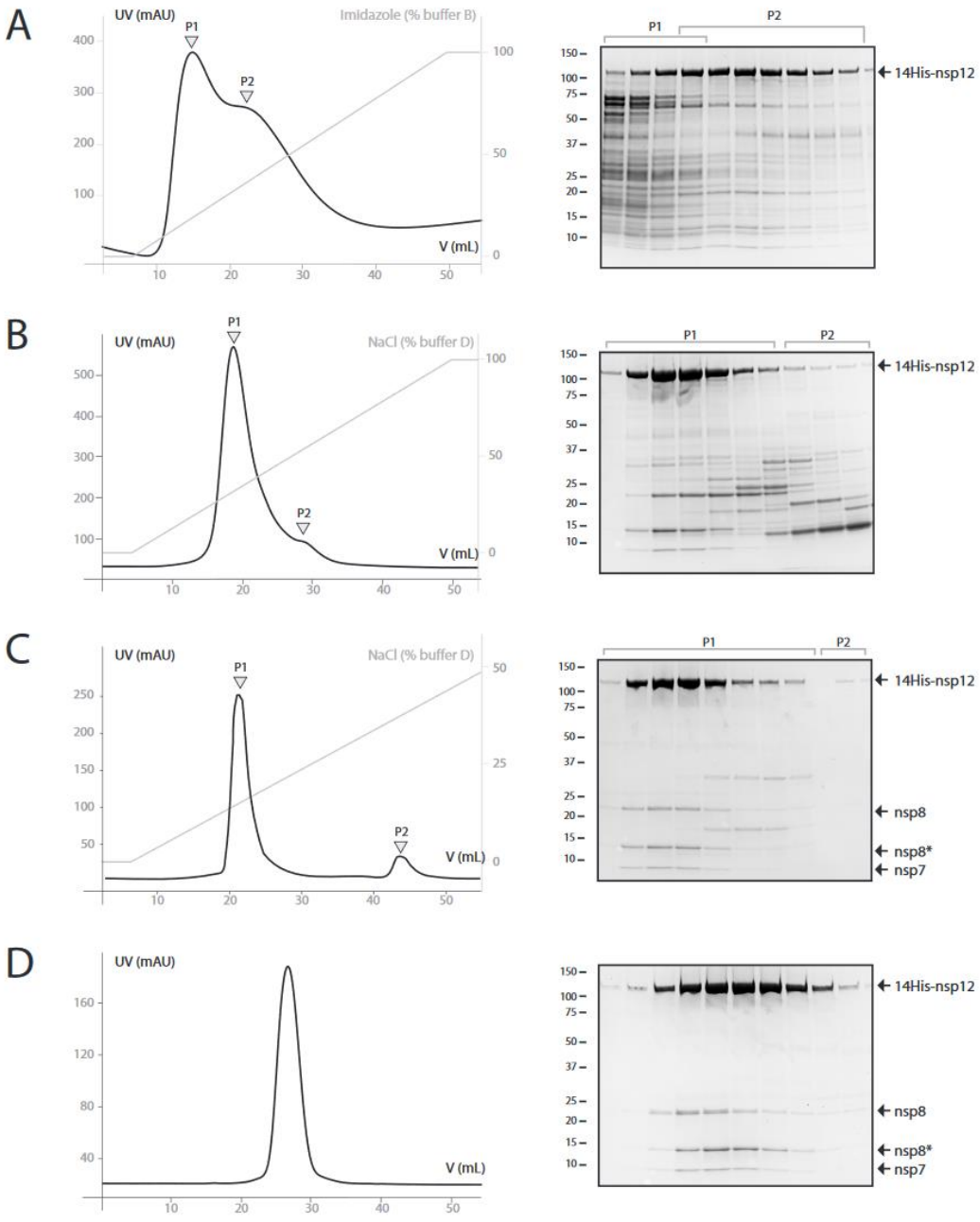


Figure 2: Purification of the recombinant SARS-CoV-2 RdRp complex

Representative elution profiles and associated SDS-PAGE (4-20%) analysis from each purification step. **(A)** Nickel affinity chromatography. Elution was performed with an imidazole gradient (10-500 mM). **(B)** Heparin affinity chromatography. Elution was performed with an NaCl gradient (50-2000 mM). **(C)** Anion exchange chromatography. Elution was performed with an NaCl gradient (50-2000 mM). **(D)** Size exclusion chromatography.

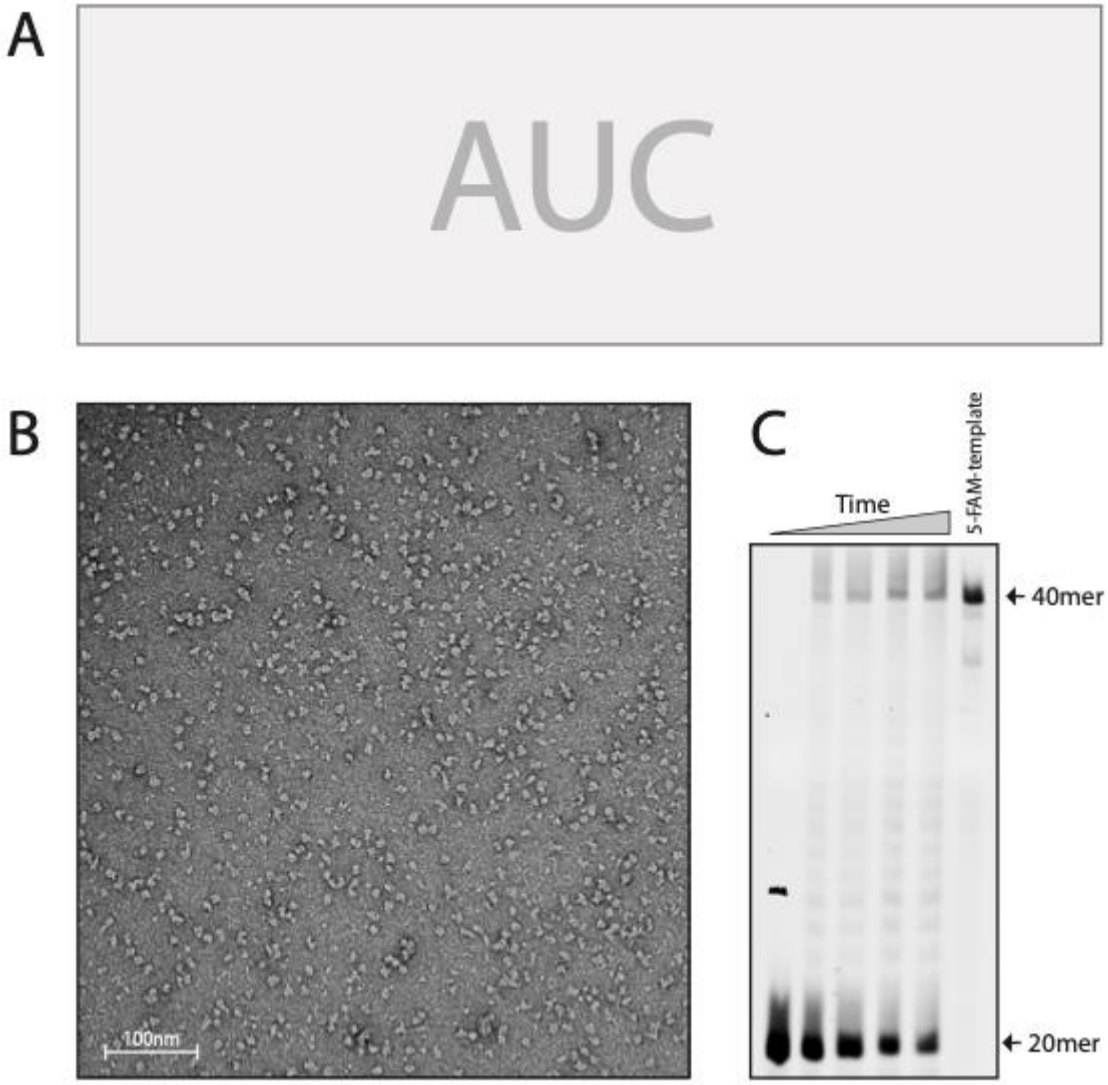


Figure 3: Biophysical and functional characterisation of the recombinant SARS-CoV-2 RdRp complex

(A) Sedimentation distribution profile of recombinant RdRp complex. The main peak shows a sedimentation coefficients 6.2 S of with a calculated molecular mass of 145 kDa, compatible with the expected (nsp12)-(nsp7)-(nsp8)-(nsp8-CTD) stoichiometry. **(B)** Representative negative-stained EM image of the recombinant RdRp complex. **(C)** RNA primer extension assay with the recombinant RdRp. The recombinant RdRp complex shows polymerase activity in vitro, extending a 20-mer primer strand labeled with a fluorescent dye at the 5′ end.

BIBLIOGRAPHY

1. D. Nelson, M. Cox, *Lehninger Principles of Biochemistry*, 7th Ed. (W. H. Freeman, 2017).

2. R. Murray, *et al.*, *Harper's Illustrated Biochemistry*, 28th Ed. (McGraw Hill Professional, 2009).

3. N. H. Acheson, *Fundamentals of Molecular Virology*, 2nd Ed. (Wiley, 2011).

4. Thomas Beddoes, *Contributions to physical and medical knowledge, principally from the West of England, collected by Thomas Beddoes, M.D.* (printed by Biggs & Cottle, for T. N. Longman and O. Rees, Paternoster-Row, London., 1799).

5. biology- definition by Encyclopedia Britannica. *https://www.britannica.com/science/biology* (August 8, 2020).

6. Biological sciences- definition by Nature. *https://www.nature.com/subjects/biological-sciences* (August 8, 2020).

7. *King James Bible, Matthew 25:40* (United Kingdom, 1611).

8. J. Orfila, Definition of intracellular pathogens. *Clinical Microbiology and Infection* **1**, S1–S2 (1996).

9. D. Raoult, *et al.*, The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).

10. Barbara E. Funnell, Gregory J. Phillips, *Plasmid Biology* (American Society of Microbiology, 2004).

11. S. Erdmann, B. Tschitschko, L. Zhong, M. J. Raftery, R. Cavicchioli, A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. *Nature Microbiology* **2**, 1446–1455 (2017).

12. E. V. Koonin, T. V. Ilyina, Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *Journal of General Virology,* **73**, 2763–2766 (1992).

13. J. P. M. Camacho, T. F. Sharbel, L. W. Beukeboom, B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **355**, 163–178 (2000).

14. H. H. Kazazian, Mobile Elements: Drivers of Genome Evolution. *Science* **303**, 1626–1632 (2004).

15. M. D. Zabel, C. Reid, A brief history of prions. *Pathog Dis* **73** (2015).

16. G. M. Cooper, The Cell: A Molecular Approach. 2nd edition (2000).

17. J. D. Marth, A unified vision of the building blocks of life. *Nature Cell Biology* **10**, 1015–1015 (2008).

18. K. R. Hagerman, P. J. Hagerman, Helix Rigidity of DNA: The Meroduplex as an Experimental Paradigm. *Journal of Molecular Biology* **260**, 207–223 (1996).

19. A. Halder, D. Data, P. P. Seelam, D. Bhattacharyya, A. Mitra, Estimating Strengths of Individual Hydrogen Bonds in RNA Base Pairs: Toward a Consensus between Different Computational Approaches. *ACS Omega* **4**, 7354–7368 (2019).

20. A. Noy, A. Pérez, F. Lankas, F. Javier Luque, M. Orozco, Relative Flexibility of DNA and RNA: a Molecular Dynamics Study. *Journal of Molecular Biology* **343**, 627–638 (2004).

21. S. Wang, E. T. Kool, Origins of the Large Differences in Stability of DNA and RNA Helixes: C-5 Methyl and 2'-Hydroxyl Effects. *Biochemistry* **34**, 4125–4132 (1995).

22. C. H. Lin, D. J. Patei, Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP. *Chemistry & Biology* **4**, 817–832 (1997).

23. A. Joachimi, A. Benz, J. S. Hartig, A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorganic & Medicinal Chemistry* **17**, 6811–6815 (2009).

24. C. R. Burke, A. Lupták, DNA synthesis from diphosphate substrates by DNA polymerases. *PNAS* **115**, 980–985 (2018).

25. A. Srivatsan, J. D. Wang, Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* **11**, 100–105 (2008).

26. R. E. Dickerson, "[5] DNA structure from A to Z" in *Methods in Enzymology*, DNA Structures Part A: Synthesis and Physical Analysis of DNA., (Academic Press, 1992), pp. 67–111.

27. J. F. Allemand, D. Bensimon, R. Lavery, V. Croquette, Stretched and overwound DNA forms a Pauling-like structure with exposed bases. *Proc Natl Acad Sci U S A* **95**, 14152–14157 (1998).

28. G. Hayashi, M. Hagihara, K. Nakatani, Application of L-DNA as a molecular tag. *Nucleic Acids Symp Ser (Oxf)* **49**, 261–262 (2005).

29. M. J. Kaplan, M. Radic, Neutrophil Extracellular Traps: Double-Edged Swords of Innate Immunity. *The Journal of Immunology* **189**, 2689–2695 (2012).

30. genetic- etymology by Online Etymology Dictionary. *https://www.etymonline.com/word/genetic* (August 12, 2020).

31. genesis- etymology by Online Etymology Dictionary. *https://www.etymonline.com/word/genesis* (August 12, 2020).

32. Richard Dawkins, *The Selfish Gene* (Oxford University Press, 1976).

33. F. H. Crick, On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).

34. Dogma- definition by Oxford Dictionary. *https://www.lexico.com/definition/dogma* (August 13, 2020).

35. Horace Freeland Judson, *The Eighth Day of Creation: The Makers of the Revolution in Biology* (1979).

36. P. N. Campbell, What mad pursuit. A personal view of scientific discovery: By Francis Crick. pp 182. Weidenfeld and Nicolson, London. 1989. £12.95 ISBN 0-297-79535-X. *Biochemical Education* **17**, 163–163 (1989).

37. James D. Watson, *et al.*, *Molecular Biology of the Gene* (Pearson, 1965).

38. H. M. Temin, S. Mizutani, Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature* **226**, 1211–1213 (1970).

39. H. M. Temin, Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol Biol Evol* **2**, 455–468 (1985).

40. retrovirus- etymology by Online Etymology Dictionary. *https://www.etymonline.com/word/retrovirus* (August 13, 2020).

41. C. W. Greider, E. H. Blackburn, Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* **43** (1985).

42. B. J. McCarthy, J. J. Holland, Denatured DNA as a direct template for in vitro protein synthesis. *PNAS* **54**, 880–886 (1965).

43. T. Uzawa, A. Yamagishi, T. Oshima, Polypeptide Synthesis Directed by DNA as a Messenger in Cell-Free Polypeptide Synthesis by Extreme Thermophiles, Thermus thermophilus HB27 and Sulfolobus tokodaii Strain 7. *J Biochem* **131**, 849–853 (2002).

44. E. V. Koonin, Does the central dogma still stand? *Biology Direct* **7**, 27 (2012).

45. F. F. Delgado, *et al.*, Intracellular Water Exchange for Measuring the Dry Mass, Water Mass and Changes in Chemical Composition of Living Cells. *PLOS ONE* **8**, e67590 (2013).

46. J. M. Goodenbour, T. Pan, Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res* **34**, 6137–6146 (2006).

47. T. Salinas-Giegé, R. Giegé, P. Giegé, tRNA Biology in Mitochondria. *International Journal of Molecular Sciences* **16**, 4518–4559 (2015).

48. E. S. Lander, *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

49. S. Li, Z. Xu, J. Sheng, tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA. *Genes* **9**, 246 (2018).

50. G. A. Khoury, R. C. Baliban, C. A. Floudas, Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports* **1**, 90 (2011).

51. M. Rother, J. A. Krzycki, Selenocysteine, Pyrrolysine, and the Unique Energy Metabolism of Methanogenic Archaea. *Archaea* **2010**, e453642 (2010).

52. E. Viñuela, M. Salas, S. Ochoa, Translation of the Genetic Message, III. Formylmethionine as Initiator of Proteins Programed by Polycistronic Messenger Rna. *PNAS* **57**, 729–734 (1967).

53. R. Bianchetti, G. Lucchini, M. L. Sartirana, Endogenous synthesis of formyl-methionine peptides in isolated mitochondria and chloroplasts. *Biochemical and Biophysical Research Communications* **42**, 97–102 (1971).

54. P. T. Wingfield, N-Terminal Methionine Processing. *Current Protocols in Protein Science* **88**, 6.14.1-6.14.3 (2017).

55. T. J. Oldfield, R. E. Hubbard, Analysis of Cα geometry in protein structures. *Proteins: Structure, Function, and Bioinformatics* **18**, 324–337 (1994).

56. V. N. Uversky, Intrinsically disordered proteins from A to Z. *The International Journal of Biochemistry & Cell Biology* **43**, 1090–1103 (2011).

57. H. Radhouani, *et al.*, Proteomic changes in extended-spectrum beta-lactamase-producing Escherichia coli strain under cefotaxime selection. *Journal of Integrated OMICS* **3**, 157-166–166 (2013).

58. Q. Ma, *et al.*, Proteomic Analysis of Ketogulonicigenium vulgare under Glutathione Reveals High Demand for Thiamin Transport and Antioxidant Protection. *PLOS ONE* **7**, e32156 (2012).

59. C. A. Hutchison, *et al.*, Design and synthesis of a minimal bacterial genome. *Science* **351** (2016).

60. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), and Nomenclature Commission of IUB (NC-IUB), Newsletter 1984. *Arch. Biochem. Biophys.*, 237–245 (1984).

61. L. S. Beese, T. A. Steitz, Structural basis for the 3′-5′ exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism. *The EMBO Journal* **10**, 25–33 (1991).

62. H. Zhu, *et al.*, Atomic structure and nonhomologous end-joining function of the polymerase component of bacterial DNA ligase D. *PNAS* **103**, 1711–1716 (2006).

63. T. A. Steitz, DNA- and RNA-dependent DNA polymerases. *Current Opinion in Structural Biology* **3**, 31–38 (1993).

64. E. E. Kim, H. W. Wyckoff, Reaction mechanism of alkaline phosphatase based on crystal structures: Two-metal ion catalysis. *Journal of Molecular Biology* **218**, 449–464 (1991).

65. T. A. Steitz, J. A. Steitz, A general two-metal-ion mechanism for catalytic RNA. *PNAS* **90**, 6498–6502 (1993).

66. M. R. Sawaya, R. Prasad, S. H. Wilson, J. Kraut, H. Pelletier, Crystal Structures of Human DNA Polymerase β Complexed with Gapped and Nicked DNA: Evidence for an Induced Fit Mechanism,. *Biochemistry* **36**, 11205–11215 (1997).

67. D. Kazlauskas, *et al.*, Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. *Journal of Molecular Biology* **430**, 737–750 (2018).

68. Y.-B. Lu, P. V. A. L. Ratnakar, B. K. Mohanty, D. Bastia, Direct physical interaction between DnaG primase and DnaB helicase of Escherichia coli is necessary for optimal synthesis of primer RNA. *PNAS* **93**, 12902–12907 (1996).

69. L. M. Iyer, E. V. Koonin, D. D. Leipe, L. Aravind, Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33**, 3875–3896 (2005).

70. H. Pan, D. B. Wigley, Structure of the zinc-binding domain of Bacillus stearothermophilus DNA primase. *Structure* **8**, 231–239 (2000).

71. L. Aravind, D. D. Leipe, E. V. Koonin, Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res* **26**, 4205–4213 (1998).

72. T. A. Guilliam, B. A. Keen, N. C. Brissett, A. J. Doherty, Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Research* **43**, 6651–6664 (2015).

73. M. P. Deutscher, A. Kornberg, Enzymatic synthesis of deoxyribonucleic acid. XXIX. Hydrolysis of deoxyribonucleic acid from the 5′ terminus by an exonuclease function of deoxyribonucleic acid polymerase. *J Biol Chem* **244**, 3029–3037 (1969).

74. H. C. Hollingsworth, N. G. Nossal, Bacteriophage T4 encodes an RNase H which removes RNA primers made by the T4 DNA replication system in vitro. *J. Biol. Chem.* **266**, 1888–1897 (1991).

75. S. García-Gómez, *et al.*, PrimPol, an Archaic Primase/Polymerase Operating in Human Cells. *Molecular Cell* **52**, 541–553 (2013).

76. B. Zhu, *et al.*, Deep-sea vent phage DNA polymerase specifically initiates DNA synthesis in the absence of primers. *PNAS* **114**, E2310–E2318 (2017).

77. M. Gabriela Kramer, S. A. Khan, M. Espinosa, Plasmid rolling circle replication: identification of the RNA polymerase-directed primer RNA and requirement for DNA polymerase I for lagging strand synthesis. *The EMBO Journal* **16**, 5784–5795 (1997).

78. K. Sakaguchi, Invertrons, a class of structurally and functionally related genetic elements that includes linear DNA plasmids, transposable elements, and genomes of adeno-type viruses. *Microbiology and Molecular Biology Reviews* **54**, 66–74 (1990).

79. M. Salas, I. Holguera, M. Redrejo-Rodríguez, M. de Vega, DNA-Binding Proteins Essential for Protein-Primed Bacteriophage Φ29 DNA Replication. *Front. Mol. Biosci.* **3** (2016).

80. M. Redrejo-Rodríguez, *et al.*, Primer-Independent DNA Synthesis by a Family B DNA Polymerase from Self-Replicating Mobile Genetic Elements. *Cell Reports* **21**, 1574–1587 (2017).

81. K. Kato, J. M. Gonalves, G. E. Houts, F. J. Bollum, Deoxynucleotide-polymerizing Enzymes of Calf Thymus Gland II. PROPERTIES OF THE TERMINAL DEOXYNUCLEOTIDYLTRANSFERASE. *J. Biol. Chem.* **242**, 2780–2789 (1967).

82. T. Kent, P. A. Mateos-Gomez, A. Sfeir, R. T. Pomerantz, Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *eLife* **5**, e13740 (2016).

83. S. Venkataraman, B. V. L. S. Prasad, R. Selvarajan, RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* **10**, 76 (2018).

84. N. Cermakian, *et al.*, On the Evolution of the Single-Subunit RNA Polymerases. *J Mol Evol* **45**, 671–681 (1997).

85. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85–98 (2011).

86. D. Forrest, K. James, Y. Yuzenkova, N. Zenkin, Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase. *Nature Communications* **8**, 15774 (2017).

87. P. Raia, M. Delarue, L. Sauguet, An updated structural classification of replicative DNA polymerases. *Biochemical Society Transactions* **47**, 239–249 (2019).

88. T. M. Ikeda, M. W. Gray, Genes and proteins of the transcriptional apparatus in mitochondria. *J Hered* **90**, 374–379 (1999).

89. H. A. M. Mönttinen, J. J. Ravantti, M. M. Poranen, Common Structural Core of Three-Dozen Residues Reveals Intersuperfamily Relationships. *Mol Biol Evol* **33**, 1697–1710 (2016).

90. L. Čuboňová, *et al.*, Archaeal DNA Polymerase D but Not DNA Polymerase B Is Required for Genome Replication in Thermococcus kodakarensis. *Journal of Bacteriology* **195**, 2322–2328 (2013).

91. Y. Ono, *et al.*, NtPolI-like1 and NtPolI-like2, Bacterial DNA Polymerase I Homologs Isolated from BY-2 Cultured Tobacco Cells, Encode DNA Polymerases Engaged in DNA Replication in Both Plastids and Mitochondria. *Plant Cell Physiol* **48**, 1679–1692 (2007).

92. S. R. Kennedy, C.-Y. Chen, M. W. Schmitt, C. N. Bower, L. A. Loeb, The Biochemistry and Fidelity of Synthesis by the Apicoplast Genome Replication DNA Polymerase Pfprex from the Malaria Parasite Plasmodium falciparum. *Journal of Molecular Biology* **410**, 27–38 (2011).

93. M. Camps, L. A. Loeb, When Pol I Goes into High Gear: Processive DNA Synthesis by Pol I in the Cell. *Cell Cycle* **3**, 114–116 (2004).

94. G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, R. Díaz-Orejas, Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62**, 434–464 (1998).

95. G. Lipps, S. Röther, C. Hart, G. Krauss, A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity. *The EMBO Journal* **22**, 2516–2525 (2003).

96. D. Kazlauskas, M. Krupovic, Č. Venclovas, The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* **44**, 4551–4564 (2016).

97. E. V. Koonin, A. E. Gorbalenya, K. M. Chumakov, Tentative identification of RNA-dependent RNA polymerases of dsRNA viruses and their relationship to positive strand RNA viral polymerases. *FEBS Letters* **252**, 42–46 (1989).

98. J. A. Jamsen, *et al.*, Time-lapse crystallography snapshots of a double-strand break repair polymerase in action. *Nature Communications* **8**, 253 (2017).

99. D. R. Stevens, S. Hammes-Schiffer, Exploring the Role of the Third Active Site Metal Ion in DNA Polymerase η with QM/MM Free Energy Simulations. *J. Am. Chem. Soc.* **140**, 8965–8969 (2018).

100. A. Bębenek, I. Ziuzia-Graczyk, Fidelity of DNA replication—a matter of proofreading. *Curr Genet* **64**, 985–996 (2018).

101. S. Ghosh, S. M. Hamdan, T. E. Cook, C. C. Richardson, Interactions of Escherichia coli Thioredoxin, the Processivity Factor, with Bacteriophage T7 DNA Polymerase and Helicase. *J. Biol. Chem.* **283**, 32077–32084 (2008).

102. L. Balakrishnan, R. A. Bambara, Flap Endonuclease 1. *Annu. Rev. Biochem.* **82**, 119–138 (2013).

103. B. A. Kelch, D. L. Makino, M. O'Donnell, J. Kuriyan, Clamp loader ATPases and the evolution of DNA replication machinery. *BMC Biology* **10**, 34 (2012).

104. T. A. Brown, C. Cecconi, A. N. Tkachuk, C. Bustamante, D. A. Clayton, Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes Dev.* **19**, 2466–2476 (2005).

105. D. Bikard, C. Loot, Z. Baharoglu, D. Mazel, Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. *Microbiol. Mol. Biol. Rev.* **74**, 570–588 (2010).

106. Publications service of Pasteur Institute (translation), On an invisible microbe antagonistic toward dysenteric bacilli: brief note by Mr. F. D'Herelle, presented by Mr. Roux. *Research in Microbiology* **158**, 553–554 (2007).

107. L. P. Villarreal, *Viruses and the Evolution of Life* (American Society of Microbiology, 2005).

108. E. J. Pritham, T. Putliwala, C. Feschotte, Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007).

109. D. Kazlauskas, A. Varsani, E. V. Koonin, M. Krupovic, Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nature Communications* **10**, 3425 (2019).

110. A. Lwoff, R. Horne, P. Tournier, A System of Viruses. *Cold Spring Harb Symp Quant Biol* **27**, 51–55 (1962).

111. A. E. Gorbalenya, *et al.*, The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology* **5**, 668–674 (2020).

112. S. Firquet, *et al.*, Survival of Enveloped and Non-Enveloped Viruses on Inanimate Surfaces. *Microbes and Environments* **30**, 140–144 (2015).

113. R. H. Symons, The intriguing viroids and virusoids: what is their information content and how did they evolve? *Mol. Plant Microbe Interact.* **4**, 111–121 (1991).

114. G. S. Diemer, K. M. Stedman, A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biology Direct* **7**, 13 (2012).

115. J. Shang, *et al.*, Cell entry mechanisms of SARS-CoV-2. *PNAS* **117**, 11727–11734 (2020).

116. R. Blumenthal, S. Durell, M. Viard, HIV Entry and Envelope Glycoprotein-mediated Fusion. *J. Biol. Chem.* **287**, 40841–40849 (2012).

117. C.-C. Pao, J. F. Speyer, Order of Injection of T7 Bacteriophage DNA. *Journal of Virology* **11**, 1024–1026 (1973).

118. P. Forterre, Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *Comptes Rendus Chimie* **14**, 392–399 (2011).

119. E. J. Walker, R. Ghildyal, Editorial: Viral Interactions with the Nucleus. *Front. Microbiol.* **8** (2017).

120. M. Schmid, T. Speiseder, T. Dobner, R. A. Gonzalez, DNA Virus Replication Compartments. *Journal of Virology* **88**, 1404–1420 (2014).

121. R. Kiro, *et al.*, Gene product 0.4 increases bacteriophage T7 competitiveness by inhibiting host cell division. *PNAS* **110**, 19549–19554 (2013).

122. Y. Shao, I.-N. Wang, Effect of Late Promoter Activity on Bacteriophage λ Fitness. *Genetics* **181**, 1467–1475 (2009).

123. H. Y. Chen, M. D. Mascio, A. S. Perelson, D. D. Ho, L. Zhang, Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. *PNAS* **104**, 19079–19084 (2007).

124. C.-C. Hu, Y.-H. Hsu, N.-S. Lin, Satellite RNAs and Satellite Viruses of Plants. *Viruses* **1**, 1325–1350 (2009).

125. B. La Scola, *et al.*, The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104 (2008).

126. M. Horie, *et al.*, Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).

127. D. DiMaio, Viruses, Masters at Downsizing. *Cell Host & Microbe* **11**, 560–561 (2012).

128. K. Limor-Waisberg, A. Carmi, A. Scherz, Y. Pilpel, I. Furman, Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. *Nucleic Acids Res* **39**, 6016–6028 (2011).

129. F. D. Giallonardo, T. E. Schlub, M. Shi, E. C. Holmes, Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *Journal of Virology* **91** (2017).

130. A. Almpanis, M. Swain, D. Gatherer, N. McEwan, Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microbial Genomics,* **4**, e000168 (2018).

131. F. W. Studier, J. J. Dunn, Organization and Expression of Bacteriophage T7 DNA. *Cold Spring Harb Symp Quant Biol* **47**, 999–1007 (1983).

132. S. K. Zavriev, M. F. Shemyakin, RNA polymerase-dependent mechanism for the stepwise T7 phage DNA transport from the virion into E. coli. *Nucleic Acids Res* **10**, 1635–1652 (1982).

133. K. E. Sloan, *et al.*, Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biology* **14**, 1138–1152 (2017).

134. P. V. Sergiev, N. A. Aleksashin, A. A. Chugunova, Y. S. Polikanov, O. A. Dontsova, Structural and evolutionary insights into ribosomal RNA methylation. *Nature Chemical Biology* **14**, 226–235 (2018).

135. P. Boccaletto, *et al.*, MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* **46**, D303–D307 (2018).

136. M. Pereira, *et al.*, Impact of tRNA Modifications and tRNA-Modifying Enzymes on Proteostasis and Human Disease. *International Journal of Molecular Sciences* **19**, 3738 (2018).

137. A. Ramanathan, G. B. Robb, S.-H. Chan, mRNA capping: biological functions and applications. *Nucleic Acids Res* **44**, 7511–7526 (2016).

138. Y. Yue, J. Liu, C. He, RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev.* **29**, 1343–1355 (2015).

139. D. Arango, *et al.*, Acetylation of Cytidine in mRNA Promotes Translation Efficiency. *Cell* **175**, 1872-1886.e24 (2018).

140. S. Kellner, The role of RNA modifications in neurological diseases- research group website. *www.cup.lmu.de/oc/kellner/research/research/*.

141. D. M. Jeziorska, *et al.*, DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *PNAS* **114**, E7526–E7535 (2017).

142. M. Tahiliani, *et al.*, Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **324**, 930–935 (2009).

143. P. Borst, R. Sabatini, Base J: Discovery, Biosynthesis, and Possible Functions. *Annu. Rev. Microbiol.* **62**, 235–251 (2008).

144. D. A. Low, N. J. Weyand, M. J. Mahan, Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence. *Infection and Immunity* **69**, 7197–7204 (2001).

145. L. Wang, *et al.*, Phosphorothioation of DNA in bacteria by dnd genes. *Nature Chemical Biology* **3**, 709–710 (2007).

146. L. Wang, *et al.*, DNA phosphorothioation is widespread and quantized in bacterial genomes. *PNAS* **108**, 2963–2968 (2011).

147. A. Kornberg, *DNA Replication* (W. H. Freeman and Company, 1980).

148. J. F. Koerner, M. S. Smith, J. M. Buchanan, Deoxycytidine Triphosphatase, an Enzyme Induced by Bacteriophage Infection. *J. Biol. Chem.* **235**, 2691–2697 (1960).

149. I. R. Lehman, E. A. Pratt, On the Structure of the Glucosylated Hydroxymethylcytosine Nucleotides of Coliphages T2, T4, and T6. *J. Biol. Chem.* **235**, 3254–3259 (1960).

150. N. Truffaut, B. Revet, M.-O. Soulie, Étude comparative des DNA de phages 2C, SP 8*, SP 82, φe, SP 01 et SP 50. *European Journal of Biochemistry* **15**, 391–400 (1970).

151. R. Tsai, I. R. Corrêa, M. Y. Xu, S. Xu, Restriction and modification of deoxyarchaeosine (dG +)-containing phage 9 g DNA. *Scientific Reports* **7**, 8348 (2017).

152. S. Ngazoa-Kakou, *et al.*, Complete Genome Sequence of Escherichia coli Siphophage BRET. *Microbiol Resour Announc* **8**, e01644-18 (2019).

153. R. Gupta, Halobacterium volcanii tRNAs. Identification of 41 tRNAs covering all amino acids, and the sequences of 33 class I tRNAs. *J. Biol. Chem.* **259**, 9461–9471 (1984).

154. G. Hutinet, *et al.*, 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nat Commun* **10**, 1–12 (2019).

155. M. D. Kirnos, I. Y. Khudyakov, N. I. Alexandrushkina, B. F. Vanyushin, 2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **270**, 369 (1977).

156. L. M. Iyer, D. Zhang, A. Maxwell Burroughs, L. Aravind, Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res* **41**, 7635–7655 (2013).

157. P. Herdewijn, P. Marlière, Toward Safe Genetically Modified Organisms through the Chemical Diversification of Nucleic Acids. *Chemistry & Biodiversity* **6**, 791–808 (2009).

158. V. B. Pinheiro, P. Holliger, The XNA world: progress towards replication and evolution of synthetic genetic polymers. *Current Opinion in Chemical Biology* **16**, 245–252 (2012).

159. H. Urata, E. Ogura, K. Shinohara, Y. Ueda, M. Akagi, Synthesis and properties of mirror-image DNA. *Nucleic Acids Res* **20**, 3325–3332 (1992).

160. S. Klußmann, A. Nolte, R. Bald, V. A. Erdmann, J. P. Fürste, Mirror-image RNA that binds D-adenosine. *Nature Biotechnology* **14**, 1112–1115 (1996).

161. E. Wyszko, *et al.*, Spiegelzymes: Sequence Specific Hydrolysis of L-RNA with Mirror Image Hammerhead Ribozymes and DNAzymes. *PLOS ONE* **8**, e54741 (2013).

162. D. Oberthür, *et al.*, Crystal structure of a mirror-image L-RNA aptamer (Spiegelmer) in complex with the natural L- protein target CCL2. *Nature Communications* **6**, 6923 (2015).

163. M. Wang, *et al.*, Mirror-Image Gene Transcription and Reverse Transcription. *Chem* **5**, 848–857 (2019).

164. V. B. Pinheiro, *et al.*, Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science* **336**, 341–344 (2012).

165. J. J. O'Brien, D. M. Campoli-Richards, Acyclovir. An Updated Review of its Antiviral Activity, Pharmacokinetic Properties and Therapeutic Efficacy. *Drugs* **37**, 233–309 (1989).

166. M. Petersen, J. Wengel, LNA: a versatile tool for therapeutics and genomics. *Trends in Biotechnology* **21**, 74–81 (2003).

167. M. B. Thayer, *et al.*, Application of Locked Nucleic Acid Oligonucleotides for siRNA Preclinical Bioanalytics. *Scientific Reports* **9**, 3566 (2019).

168. D. A. Malyshev, *et al.*, A semi-synthetic organism with an expanded genetic alphabet. *Nature* **509**, 385–388 (2014).

169. Y. Zhang, *et al.*, A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proceedings of the National Academy of Sciences* **114**, 1317–1322 (2017).

170. A. Marx, K. Betz, The Structural Basis for Processing of Unnatural Base Pairs by DNA Polymerases. *Chemistry – A European Journal* **26**, 3446–3463 (2020).

171. S. Hoshika, *et al.*, Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* **363**, 884–887 (2019).

172. I. Ya. Khudyakov, M. D. Kirnos, N. I. Alexandrushkina, B. F. Vanyushin, Cyanophage S-2L contains DNA with 2,6-diaminopurine substituted for adenine. *Virology* **88**, 8–18 (1978).

173. M. Cristofalo, *et al.*, Nanomechanics of Diaminopurine-Substituted DNA. *Biophysical Journal* **116**, 760–771 (2019).

174. D. C. Ward, E. Reich, L. Stryer, Fluorescence Studies of Nucleotides and Polynucleotides I. FORMYCIN, 2-AMINOPURINE RIBOSIDE, 2,6-DIAMINOPURINE RIBOSIDE, AND THEIR DERIVATIVES. *J. Biol. Chem.* **244**, 1228–1237 (1969).

175. C. Santhosh, P. C. Mishra, Electronic spectra of 2-aminopurine and 2,6-diaminopurine: phototautomerism and fluorescence reabsorption. *Spectrochimica Acta Part A: Molecular Spectroscopy* **47**, 1685–1693 (1991).

176. M. Szekeres, A. V. Matveyev, Cleavage and sequence recognition of 2,6-diaminopurine-containing DNA by site-specific endonucleases. *FEBS Letters* **222**, 89–94 (1987).

177. A. Solís-Sánchez, *et al.*, Genetic characterization of ØVC8 lytic phage for Vibrio cholerae O1. *Virology Journal* **13**, 47 (2016).

178. S. Pennell, *et al.*, FAN1 Activity on Asymmetric Repair Intermediates Is Mediated by an Atypical Monomeric Virus-type Replication-Repair Nuclease Domain. *Cell Reports* **8**, 84–93 (2014).

179. P. Weber, *et al.*, High-Throughput Crystallization Pipeline at the Crystallography Core Facility of the Institut Pasteur. *Molecules* **24**, 4451 (2019).

180. P. Legrand, *XDS Made Easier (2017) GitHub repository*.

181. D. Liebschner, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst D* **75**, 861–877 (2019).

182. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Cryst D* **66**, 486–501 (2010).

183. M. Delarue, Y.-H. Sanejouand, Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model. *Journal of Molecular Biology* **320**, 1011–1024 (2002).

184. E. Lindahl, C. Azuara, P. Koehl, M. Delarue, NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res* **34**, W52–W56 (2006).

185. T. A. Steitz, DNA Polymerases: Structural Diversity and Common Mechanisms. *J. Biol. Chem.* **274**, 17395–17398 (1999).

186. E. Loh, L. A. Loeb, Mutability of DNA polymerase I: Implications for the creation of mutant DNA polymerases. *DNA Repair* **4**, 1390–1398 (2005).

187. Y. Li, S. Korolev, G. Waksman, Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *The EMBO Journal* **17**, 7514–7525 (1998).

188. V. Derbyshire, N. D. Grindley, C. M. Joyce, The 3′-5′ exonuclease of DNA polymerase I of Escherichia coli: contribution of each amino acid at the active site to the reaction. *The EMBO Journal* **10**, 17–24 (1991).

189. M. H. Buckstein, J. He, H. Rubin, Characterization of Nucleotide Pools as a Function of Physiological State in Escherichia coli. *Journal of Bacteriology* **190**, 718–726 (2008).

190. H.-J. Kim, S. A. Benner, Prebiotic stereoselective synthesis of purine and noncanonical pyrimidine nucleotide from nucleobases and phosphorylated carbohydrates. *PNAS* **114**, 11315–11320 (2017).

191. J. P. Schrum, A. Ricardo, M. Krishnamurthy, J. C. Blain, J. W. Szostak, Efficient and Rapid Template-Directed Nucleic Acid Copying Using 2′-Amino-2′,3′-dideoxyribonucleoside–5′-Phosphorimidazolide Monomers. *J. Am. Chem. Soc.* **131**, 14560–14570 (2009).

192. E. Nudler, RNA Polymerase Backtracking in Gene Regulation and Genome Instability. *Cell* **149**, 1438–1445 (2012).

193. A. Singh, *et al.*, Excessive excision of correct nucleotides during DNA synthesis explained by replication hurdles. *The EMBO Journal* **39**, e103367 (2020).

194. D. Kazlauskas, M. Krupovic, J. Guglielmini, P. Forterre, Č. Venclovas, Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res* **48**, 10142–10156 (2020).

195. J. E. Lohr, F. Chen, R. T. Hill, Genomic Analysis of Bacteriophage ΦJL001: Insights into Its Interaction with a Sponge-Associated Alpha-Proteobacterium. *Appl. Environ. Microbiol.* **71**, 1598–1609 (2005).

196. E. E. Kulikov, *et al.*, Genomic Sequencing and Biological Characteristics of a Novel Escherichia Coli Bacteriophage 9g, a Putative Representative of a New Siphoviridae Genus. *Viruses* **6**, 5077–5092 (2014).

197. M. Tsai, *et al.*, Substrate and Product Complexes of Escherichia coli Adenylosuccinate Lyase Provide New Insights into the Enzymatic Mechanism. *Journal of Molecular Biology* **370**, 541–554 (2007).

198. K. D. Tardif, J. Horowitz, Functional group recognition at the aminoacylation and editing sites of E. coli valyl-tRNA synthetase. *RNA* **10**, 493–503 (2004).

199. C. Trzaska, *et al.*, 2,6-Diaminopurine as a highly potent corrector of UGA nonsense mutations. *Nature Communications* **11**, 1509 (2020).

200. Feynman's blackboard at the time of his death, Caltech archives (available at archives.caltech.edu/pictures/1.10-29.jpg).

201. P. F. Crain, Preparation and enzymatic hydrolysis of DNA and RNA for mass spectrometry. *Methods in Enzymology* **193**, 782–790 (1990).

202. *Serial Cloner 2.6 (available at serialbasics.free.fr/Serial_Cloner.html)*.

203. M. P. Callahan, *et al.*, Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *PNAS* **108**, 13995–13998 (2011).

204. H. F. Schmidt, E. G. Sakowski, S. J. Williamson, S. W. Polson, Ke. Wommack, Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *The ISME Journal* **8**, 103–114 (2014).

205. J. Filée, P. Forterre, T. Sen-Lin, J. Laurent, Evolution of DNA Polymerase Families: Evidences for Multiple Gene Exchange Between Cellular and Viral Proteins. *J Mol Evol* **54**, 763–773 (2002).

206. T. W. Schoenfeld, *et al.*, Lateral Gene Transfer of Family A DNA Polymerases between Thermophilic Viruses, Aquificae, and Apicomplexa. *Mol Biol Evol* **30**, 1653–1664 (2013).

207. S. Lu, *et al.*, CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* **48**, D265–D268 (2020).

208. T. U. Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).

209. F. Madeira, *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47**, W636–W641 (2019).

210. T. Frickey, A. Lupas, CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).

211. K. Takata, *et al.*, Analysis of DNA polymerase ν function in meiotic recombination, immunoglobulin class-switching, and DNA damage tolerance. *PLOS Genetics* **13**, e1006818 (2017).

212. K. Makiela-Dzbenska, *et al.*, Role of Escherichia coli DNA polymerase I in chromosomal DNA replication fidelity. *Molecular Microbiology* **74**, 1114–1127 (2009).

213. R. Inoue, *et al.*, Genetic identification of two distinct DNA polymerases, DnaE and PolC, that are essential for chromosomal DNA replication in Staphylococcus aureus. *Mol Gen Genomics* **266**, 564–571 (2001).