UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA

Programa de Pós-graduação em Ciências Biológicas (Bioquímica)

SANTIAGO JUSTO ARÉVALO

Estudos na especificidade enzimática de diguanilato ciclases e na produção de novos segundos mensageiros

Studies on the enzymatic specificity of diguanylate cyclases domains and the production of new second messengers

Versão corrigida

São Paulo

Data do Depósito na SPG:

27/04/2022

SANTIAGO JUSTO ARÉVALO

Estudos na especificidade enzimática de diguanilato ciclases e na produção de novos segundos mensageiros

Studies on the enzymatic specificity of diguanylate cyclases domains and the production of new second messengers

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioquímica)

Orientador: Prof. Dr. Shaker Chuck Farah

São Paulo

2022

Universidade de São Paulo
Instituto de Química

"Estudos na especificidade enzimática de diguanilato ciclases e na produção de novos segundos mensageiros"

SANTIAGO JUSTO AREVALO

Tese de Doutorado submetida ao Instituto de Química da Universidade de São Paulo como parte dos requisitos necessários à obtenção do grau de Doutor em Ciências obtido no Programa Ciências Biológicas (Bioquímica) - Área de Concentração: Bioquímica.

____
Prof. Dr. Shaker Chuck Farah
(Orientador e Presidente)

APROVADO(A) POR:

____
Prof. Dr. Sandro Roberto Marana
IQ - USP

____
Prof. Dr. Carlos Henrique Inacio Ramos
IQ - UNICAMP

____
Profa. Dra. Ljubica Tasic
IQ - UNICAMP

SÃO PAULO
15 de julho de 2022

Ficha Catalográfica elaborada eletronicamente pelo autor, utilizando o
programa desenvolvido pela Seção Técnica de Informática do ICMC/USP e
adaptado para a Divisão de Biblioteca e Documentação do Conjunto das Químicas da USP

Bibliotecária responsável pela orientação de catalogação da publicação:
Marlene Aparecida Vieira - CRB - 8/5562

Justo Arévalo, Santiago
J96e Estudos na especificidade enzimática de
diguanilato ciclases e na produção de novos segundos
mensageiros / Santiago Justo Arévalo. - São Paulo,
2022.
210 p.

Tese (doutorado) - Instituto de Química da
Universidade de São Paulo. Departamento de
Bioquímica.
Orientador: Farah, Shaker Chuck

1. Biologia molecular. 2. Microbiologia. 3.
Enzimologia. 4. Biologia Estrutural. I. T. II.
Farah, Shaker Chuck, orientador.

A todos os que como estrelas ainda nos alumbram a pesar que sua luz já está extinguida.

ACKNOWLEDGEMENTS

Aos meus pais por sempre manterem o seu apoio e confiança em mim, e por terem sido uma grande parte da minha inspiração para iniciar, continuar e terminar esta pós-graduação.

Aos meus irmãos por me ensinarem com suas ações, me ajudarem a relaxar, me distrair e clarear minha mente nas minhas viagens para visitar minha casa.

A todos os meus familiares em geral, simplesmente por estarem ali, cada um com seu jeito de ser inspira algo diferente em mim.

À Daniela Zapata Sifuentes por sua companhia em todos os momentos, nossas conversas que impulsionam meu desejo de aprender mais.

Ao Prof. Shaker Chuck Farah por sua orientação, seus ensinamentos, por compartilhar suas hipóteses e ideias e acima de tudo pela liberdade que me deu para fluir e testar minhas próprias ideias.

Aos alunos de iniciação científica (Cindy, Carla, Marlon, Claudia, Maria, Tania, Janny, Andrea, Michella, Daniela, Adriana, Izabo, Matheus) com quem pude compartilhar durante várias etapas do meu doutorado, aprendi muito com cada um deles.

Aos membros do Laboratório do Prof. Chuck que, com suas conversas, apresentações e discussões, me permitiram perceber o ambiente acadêmico de uma forma diferente.

Ao William Cenens por me ajudar nas análises microscópicas dos biossensores.

Ao Bruno Matsuyama por me ajudar a obter modelos estruturais de proteínas a partir de dados cristalográficos.

Ao Prof. Ray Owens por me convidar e me dar a oportunidade de trabalhar na Oxford Protein Production Facility.

Ao Prof. João Renato e Deyvid Amgarten por me incluírem em sua equipe no Laboratório Clínico do Hospital Albert Einstein em São Paulo, onde aprendi muito sobre análise de sequência viral.

À Profa. Aline e o Prof. João Setúbal por nos orientar no projeto relacionado às análises genômicas e bioquímicas de *Bacillus* degradadores de cianeto.

À Profa. Flavia Meoti e sua aluna Railmara por me ensinarem como usar e analisar resultados de HPLC.

Ao Prof. Mauro e ao Prof. Alcides por suas contribuições ao projeto relacionado ao *Bacillus* degradadores de cianeto e aos procedimentos que realizaram para que os alunos de iniciação científica da Universidade Ricardo Palma (Lima - Peru) pudessem vir à USP para ter uma experiência acadêmica e colaborar nos projetos que realizei.

Aos administradores do cluster de informática da Universidade Ricardo Palma (Lima - Peru), Gustavo Abarca e Roxana Mier, e ao administrador do Instituto de Pesquisas da Amazônia Peruana (Loreto - Peru), Rodolfo Cárdenas, pela assistência na instalação de software e gerenciamento de cluster.

À Vania Bueno e Giovanna de Freitas, por me ensinarem espectrometria de massas e realizarem as centenas de corridas que foram feitas durante meu doutorado.

Aos trabalhadores da Central Analítico (IQ-USP) em geral, pelo trabalho que realizam para facilitar nossas tarefas com equipamentos que exigem experiência para manusear.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) processo nº 2015/13318-4 e o Conselho Nacional de Pesquisa (CNPq) pelo auxílio financeiro.

RESUMO

Santiago Justo Arévalo. **Estudos na especificidade enzimática de diguanilato ciclases e na produção de novos segundos mensageiros.** 2022. 210p. Tese (Doutorado) – Programa de Pós-Graduação em Ciências (Bioquímica). Instituto de Química, Universidade de São Paulo, São Paulo.

O presente trabalho está dividido em três capítulos sobre linhas de pesquisa diferentes desenvolvidas pelo autor durante o período de doutorado

No primeiro capítulo, são apresentados estudos relacionados ao reconhecimento estrutural de substratos e análise enzimática de domínios GGDEF com atividade diguanilato ciclase (EC 2.7.7.65). As proteínas contendo domínios GGDEF estão relacionados à produção enzimática do segundo mensageiro c-di-GMP, a partir de duas moléculas de GTP, em procariotos. Esta molécula está principalmente envolvida na transição entre os estilos de vida móveis e sésseis, bem como vários outros fenótipos. Redundância e diversidade de sequências de domínio GGDEF aumentam a possibilidade de que outras funções enzimáticas ainda possam ser descobertas. Para testar esta hipótese, i) o efeito de mutações pontuais na estrutura e atividade enzimática dos domínios GGDEF é analisado, ii) a especificidade enzimática de domínios GGDEF de enzimas diferentes também é testada e iii) quando produtos não canônicos são detectados, modelos enzimáticos são estudados para entender sua produção preferencial. Como resultados mais importantes, sete mutantes do PleD (uma proteína contendo GGDEF) foram construídos e a estrutura cristalográfica de dois delas foi resolvida, mostrando que é improvável que eles liguem à porção guanina em seu sítio ativo. Além disso, cinco mutantes da proteína XAC0610 de *Xanthomonas citri* foram construídos e sua capacidade de usar ATP ou GTP como substrato foi avaliada. Nenhum desses mutantes foi capaz de usar ATP como substrato. Finalmente, sete outras proteínas contendo GGDEF foram purificadas e sua especificidade enzimática foi avaliada com vários trifosfatos de nucleotídeos. Uma enzima promíscua chamada GSU1658 mostrou produzir c-di-GMP, c-di-AMP, c-di-IMP, c-di-2´dGMP, c-GAMP, c-GIMP e c-AIMP. Curiosamente, o XAC0610 foi capaz de reconhecer 2´dGTP como substrato. A análise da cinética enzimática de XAC0610 na presença de 2´dGTP e GTP mostrou a formação preferencial do produto linear híbrido pppGp2´dG.

O segundo capítulo aborda estudos sobre o metabolismo do cianeto em *Bacillus* com foco na cianeto dihidratase de *Bacillus safensis*. O cianeto é amplamente utilizado nas indústrias devido à sua alta afinidade com os metais. Esta mesma capacidade confere toxicidade potente a este composto. Assim, as indústrias têm que reduzir a concentração de cianeto das águas residuais antes de sua disposição final. Métodos físicos, químicos e biológicos têm sido desenvolvidos para atingir esse objetivo, mas o conhecimento sobre as vias metabólicas e a biologia das enzimas envolvidas na degradação do cianeto ainda é escasso. Aqui, é descrito o isolamento de uma cepa de *Bacillus safensis* de rejeitos de minas no Peru. A classificação desta cepa foi feita através de uma análise comparativa de 132 “core genomes” de cepas do grupo de *Bacillus* *pumilus*. Em seguida, determinamos que uma cianeto dihidratase (CynD, EC 3.5.5.1) codificada no genoma da cepa isolada era provavelmente a enzima responsável pela degradação do cianeto. A confirmação da atividade degradante de cianeto de CynD desta cepa foi feita por clonagem, expressão e purificação da enzima e realização de caracterização enzimática. O CynD desta cepa é ativo até pH 9 e os padrões de oligomerização analisados por SEC-MALS mostraram que a enzima forma longas estruturas helicoidais em pH 8 e estruturas menores enquanto o pH aumenta. Finalmente, foi demonstrado que a expressão de CynD é fortemente induzida na presença de cianeto.

Os últimos dois anos do doutorado foram realizados no contexto da pandemia COVID-19. Vários laboratórios se dedicaram a gerar conhecimento para ajudar no combate à pandemia. Nesta situação e graças à grande quantidade de dados genômicos disponíveis publicamente, estudos sobre a dinâmica das mutações do SARS-CoV-2 foram realizados. No primeiro ano da pandemia, a classificação genômica de 171.461 genomas mostrou a presença de cinco haplótipos principais com base em nove mutações. A distribuição mundial e a mudança de frequência desses haplótipos foram analisadas cuidadosamente. Todos os haplótipos foram identificados nas seis regiões analisadas (América do Sul, América do Norte, Europa, Ásia, África e Oceania); no entanto, a frequência de cada um deles foi diferente em cada uma dessas regiões. Em 30 de setembro de 2020, o haplótipo 3 (ou unidade taxonômica operacional 3, OTU_3) era o mais prevalente em quatro regiões (América do Sul, Ásia, África e Oceania). OTU_5 foi o mais prevalente na América do Norte e OTU_2 na Europa. A dinâmica temporal dos haplótipos mostrou que OTU_1 parece perto da extinção após 8 meses de pandemia (novembro de 2020). Outros OTUs ainda estão presentes em diferentes frequências em todo o mundo, mesmo atualmente gerando novas variantes. Com base em sua dinâmica temporal, um esquema de classificação de 115 mutações SARS-CoV-2 identificadas a partir de 1.058.020 genomas SARS-COV-2 também foi feito. Três tipos de dinâmica temporal de mutações foram identificados: i) Mutações de alta frequência, ii) mutações de média frequência e iii) mutações de baixa frequência. Finalmente, foi analisada a correlação do número de reprodução efetiva (Rt) do SARS-CoV-2 que contém a mutação de alta frequência N501Y com o nível de medidas de controle, mostrando que seu Rt está negativamente correlacionado com o nível de medidas de controle em oito dos nove países analisados. Esta correlação negativa foi semelhante quando foi analisado o Rt de SARS-CoV-2 sem a mutação N501Y. Assim, as medidas de controle provavelmente diminuirão o Rt de SARS-CoV-2 “tipo selvagem” e N501Y.

Palabras chave: Dominios GGDEF, dinucleotideos cíclicos, *Xanthomonas citri*, PleD, XAC0610, CynD, *Bacillus safensis*, Cianeto, Core genomes, SARS-CoV-2, mutações.

ABSTRACT

Santiago Justo Arévalo. **Studies on the enzymatic specificity of diguanylate cyclases domains and the production of new second messengers.** 2022. 210p. PhD Thesis – Graduate Program in Sciences (Biochemistry). Instituto de Química, Universidade de São Paulo, São Paulo.

The work presented in this thesis is divided into three chapters in which the author presents the results of three different lines of research that were carried out during his PhD.

In the first chapter, studies on substrate recognition and enzymatic activity of GGDEF domains are presented. Many proteins containing GGDEF domains are diguanylate cyclases (DGCs, EC 2.7.7.65), enzymes that catalyze the conversion of 2 GTP molecules into the second messenger c-di-GMP in prokaryotes. This molecule is primarily implicated in the transition between motile and sessile lifestyles, as well several other phenotypes. Redundancy and diversity of GGDEF domain sequences in many bacterial genomes raises the possibility that other enzymatic functions may yet be discovered. To test this hypothesis, i) the effect of point mutations on the structure and enzymatic activity of GGDEF domains is analyzed, ii) the enzymatic specificity of wild-type GGDEF domains from different proteins is also tested, and iii) when non-canonical products are detected, enzymatic models are studied to understand its preferential production. The principal results obtained from these studies are as follows. Seven mutants of the DGC PleD (a GGDEF containing-protein from *Caulobacter crescentus*) were constructed and the crystallographic structure of two of them was solved, showing that they are unlikely to bind the guanine moiety in its active site. Additionally, five mutants of XAC0610, another DGC from *Xanthomonas citr*, were constructed and their substrate specificities were evaluated. None of those mutants were able to use ATP as a substrate. Finally, seven different GGDEF domain-containing DGCs from different sources were expressed and purified and their enzymatic specificities were tested with several nucleotide triphosphates. One enzyme, GSU1658 from *Geobacter sulfurreducens* was particularly promiscuous and shown to produce c-di-GMP, c-di-AMP, c-di-IMP, c-di-2´dGMP, c-GAMP, c-GIMP, and c-AIMP. Interestingly, XAC0610 was able to recognize 2´dGTP as substrate. Analysis of enzyme kinetics of XAC0610 in presence of 2´dGTP and/or GTP showed the preferential formation of the hybrid linear product pppGp2´dG.

The second chapter present studies on cyanide metabolism in *Bacillus* with focus on the cyanide dihydratase of *Bacillus safensis*. Cyanide is widely used in industries due to its high affinity for metals. This same ability confers potent toxicity to this compound. Thus, industries must reduce the cyanide concentration from wastewater before its final disposal. Physical, chemical, and biological methods have been developed to achieve this goal, but knowledge about metabolic pathways and the biology of enzymes involved in cyanide degradation is still scarce. Here, the isolation of a *Bacillus safensis* strain from mine tailings in Peru is described. Classification of this strain was done through a comparative analysis of 132 core genomes of strains from the *Bacillus pumilus* group.

Sequence analysis determined that a cyanide dihydratase (CynD, EC 3.5.5.1)) encoded in the genome of the isolated strain was likely the enzyme responsible for cyanide degradation. Confirmation of the cyanide degrading activity of CynD from this strain was achieved by cloning, expression and purification of the enzyme and its enzymatic characterization. CynD from this strain was active up to pH 9 and oligomerization patterns analyzed by SEC-MALS and electron microscopy showed that the enzyme forms large helical structures at pH 8 and smaller structures at higher pHs. Finally, we show that CynD expression is strongly induced in the presence of cyanide.

The last two years of graduate studies were carried out in the context of the COVID-19 pandemic. Thanks to the large amount of publicly available genomic data, we were able to carry out studies on the worldwide dynamics of the spread of SARS-CoV-2 mutants forms. In the first year of the pandemic, genomic classification of 171,461 genomes showed the presence of five major haplotypes based on nine mutations. The worldwide distribution and the temporal evolution of frequency of these haplotypes was carefully analyzed. All the haplotypes were identified in the six regions analyzed (South America, North America, Europe, Asia, Africa, and Oceania); however, the frequency of each of them was different in each of these regions. As of September 30, 2020, haplotype 3 (or operational taxonomic unit 3, OTU_3) was the most prevalent in four regions (South America, Asia, Africa, and Oceania). OTU_5 was the most prevalent in North America and OTU_2 in Europe. Temporal dynamics of the haplotypes showed that OTU_1 became nearly extinct after 8 months of pandemic (November 2020). Other OTUs are still present in different frequencies all around the world, while currently generating new variants. Based on their temporal dynamics, a classification scheme of 115 SARS-CoV-2 mutations identified from 1,058,020 SARS-COV-2 genomes was also performed. Three types of temporal dynamics of mutations were identified: i) High-Frequency mutations are characterized by a rapid increase in frequency upon its appearance, ii) medium and iii) low-frequency mutations maintain mid or low-frequencies for several months and can be region-specific. Finally, we performed a correlation analysis of the effective reproduction number (Rt) of SARS-CoV-2 harboring the high-frequency mutation N501Y with the level of control measures adopted in specific jurisdictions. We show that Rt is negatively correlated with the level of control measures in eight of the nine countries analyzed. This negative correlation was similar when we analyzed the Rt of SARS-CoV-2 not-harboring N501Y. Thus, the control measures likely diminish the Rt of both SARS-CoV-2 “wild-type” and N501Y.

Key words: GGDEF domains, cylic dinucleotides, *Xanthomonas citri*, PleD, XAC0610, CynD, *Bacillus safensis*, Cyanide, Core genomes, SARS-CoV-2, mutations.

INDEX

CHAPTER 1. ENZYMATIC AND STRUCTURAL STUDIES ON GGDEF DOMAINS OF DIGUANYLATE CYCLASES

1.1.- INTRODUCTION:

1.1.1.- Cyclic nucleotides:

The cyclic nucleotides are very important signaling molecules both in eukaryotes and in prokaryotes, due to their involvement in several fundamental pathways mainly in response to changes in environmental factors. There are two main groups of these molecules: the cyclic nucleotides and the cyclic dinucleotides (Fig. 1).

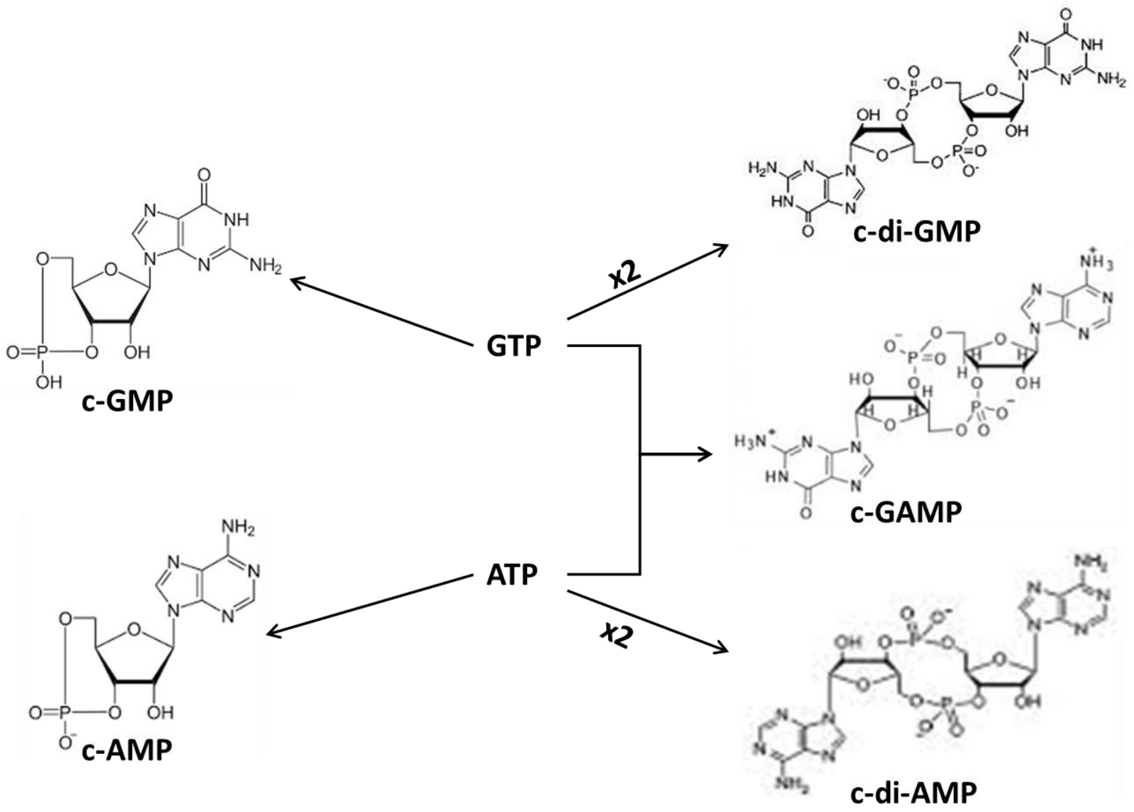


Figure 1: Second messenger cyclic nucleotides. At left the group of the cyclic nucleotides: cAMP and cGMP produced from an ATP or GTP molecule respectively. At Right the cyclic dinucleotides group produced from two molecules of ATP or GTP in the case of c-di-GMP and c-di-AMP, and for one molecule of each one in the case of c-GAMP.

The first group is very well characterized and consists of two important molecules: Cyclic adenosine monophosphate (c-AMP) and cyclic guanosine monophosphate (c-GMP). These are produced from an ATP molecule and a GTP molecule respectively by adenylyl cyclase (AC) and guanylyl cyclase (GC) domains respectively. The AC and GC domains are important in the activation of kinases and in the regulation of ion channels in a wide range of cells (Tucker et al., 1998). The second group, the cyclic dinucleotides, is composed of three, relatively recently discovered, important molecules: cyclic diadenosine monophosphate (c-di-AMP), cyclic diguanosine monophosphate (c-di-GMP) and cyclic guanosine-adenosine monophosphate (c-GAMP).

1.1.2.- c-di-AMP:

The c-di-AMP molecule has fundamental roles in microbial growth and physiology in Gram-positive bacteria (Fig. 2A), specifically in the stress response, antibiotic resistance, cellular morphology, and virulence (Bai et al., 2012; Corrigan et al., 2011, 2013; Corrigan & Gründling, 2013; Sureka et al., 2014; Witte et al., 2008). It has also been implicated in the activation of the immune response of the host, activating the production of IFN-b (Yang et al., 2014).

The concentration of c-di-AMP is controlled by the diadenylate cyclases (DACs), that produce c-di-AMP from two molecules of ATP or ADP (Bai et al., 2012), and by the c-di-AMP phosphodiesterase, that degrades c-di-AMP to pApA or AMP (Bai et al., 2013; Corrigan et al., 2011) (Fig. 2A).

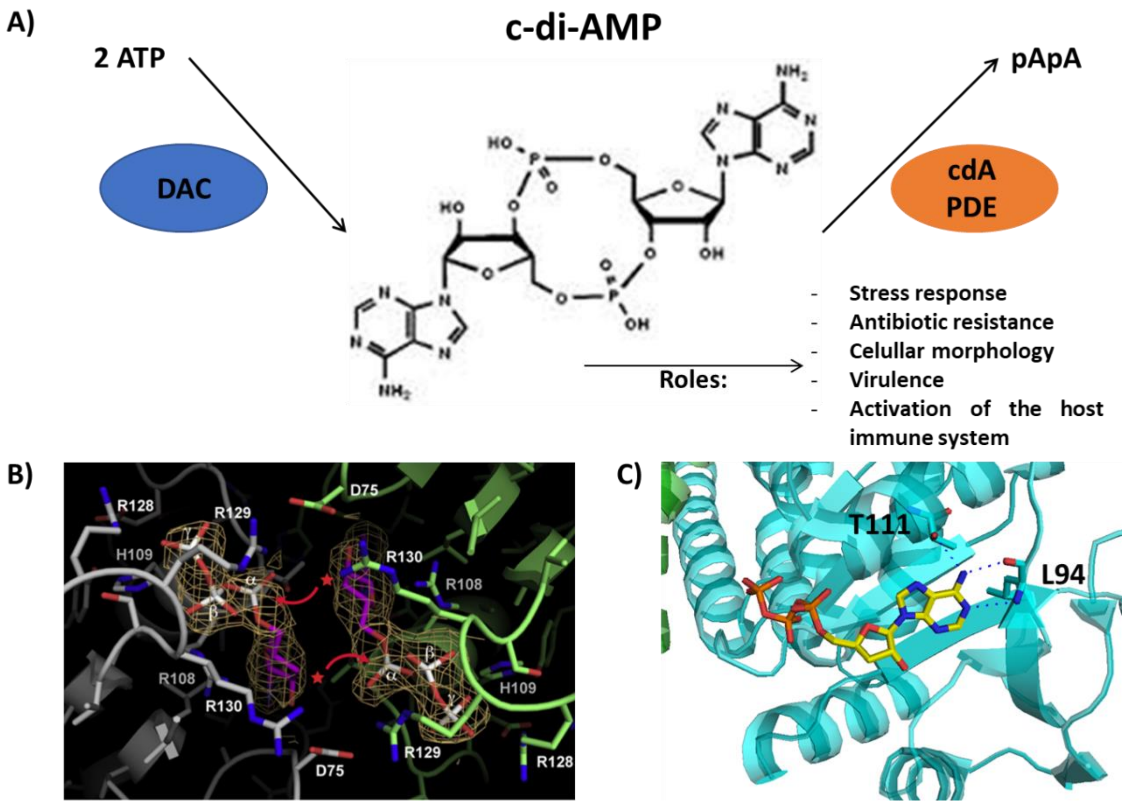


Figure 2: c-di-AMP metabolism. A) c-di-AMP is produced by Diadenylate cyclase proteins (DACs) and degraded to pApA by c-di-AMP phosphodiesterase (cdA PDE), this cyclic nucleotide has several demonstraded roles. B) and C) Crystallography structure of DisA in complex with non-reactive ATP (3´-deoxy ATP), B) showing the residues implicated in the cyclization reaction (Witte G et al. 2008) and C) showing the residues implicated in the specific recognition of the nucleotide base adenine.

One well characterized diadenylate cyclase protein is DisA (DNA integrity scanning protein A) of *Bacillus subtilis* and its homolog in *Thermotoga maritima*. This octameric protein senses the presence of a DNA double-strand breaks and monitor proper genome replication and genome integrity during the sporulation process. Witte et al. (2008) crystallized DisA of *T. maritima* in a complex with 3´-deoxy ATP that lacks the nucleophilic 3´-OH group. They proposed that DisA can binds one ATP in each of the eight monomers and thus the octamer has four different active sites where it can produce cdiAMP. They also showed that several conserved acidic and basic side chains in the RHR and DGA motif of DisA are possibly implicated in the cyclase reaction, including R128, R129, H109, R1308, R130 and D75 (Fig. 2B). They also show that DisA of both *B. subtilis* and *T. maritima* are very specific for ATP. Reviewing the DisA structure in complex with non-hydrolyzable ATP (3´-deoxy ATP) (PDB:3C23), two residues Interact with the N6 amine group of the nucleotide base: the main chain carbonyl group of L94 and the OH group of the T111 side chain (Fig. 2C).

1.1.3.- C-GAMP:

Another signaling molecule, c-GAMP (cyclic guanosine adenosine monophosphate) (Fig 3), is formed from the condensation of an ATP molecule and a GTP molecule. c-GAMP has been found as a part of signaling cascades in eukaryotes (mammals) (Sun et al., 2013) and in prokaryotes, in *Vibrio* (Davies et al., 2012) and *Geobacter* (Hallberg et al., 2016, 2019).

In mammals, c-GAMP is produced by the cGAS protein (Cyclic Guanine Adenine Synthetase), a monomeric enzyme with capacity to simultaneously bind ATP and GTP. cGAMP produced by cGAS has one 2´-5´ phosphodiester link and another 3´-5´ phosphodiester link (Ablasser et al., 2013). This molecule participates in the STING signaling pathway that activates the production of IFN-B in response to cytoplasmic DNA (Sun et al., 2013)

In the prokaryote *Vibrio cholerae*, the causative agent of cholera, this molecule is produced by the DncV protein (Dinucleotide Cyclase of Vibrio) (Davies et al., 2012). Unlike the mammalian c-GAMP, the bacterial molecule has two 3´-5´ phosphodiester linkages (Diner et al., 2013) and is involved in the control of folate metabolism in *Vibrio* *cholerae* (Zhu et al., 2014) and in a recently discovered mechanism of bacterial protection against phages in different prokaryotes such as: *Vibrio cholerae*, *Thioalkalimicrobium aerophilum*, *Dechloromonas agitate*, *Yersinia kristensenii*, *Burkholderia sp.*, *E. coli* (Cohen et al., 2019).

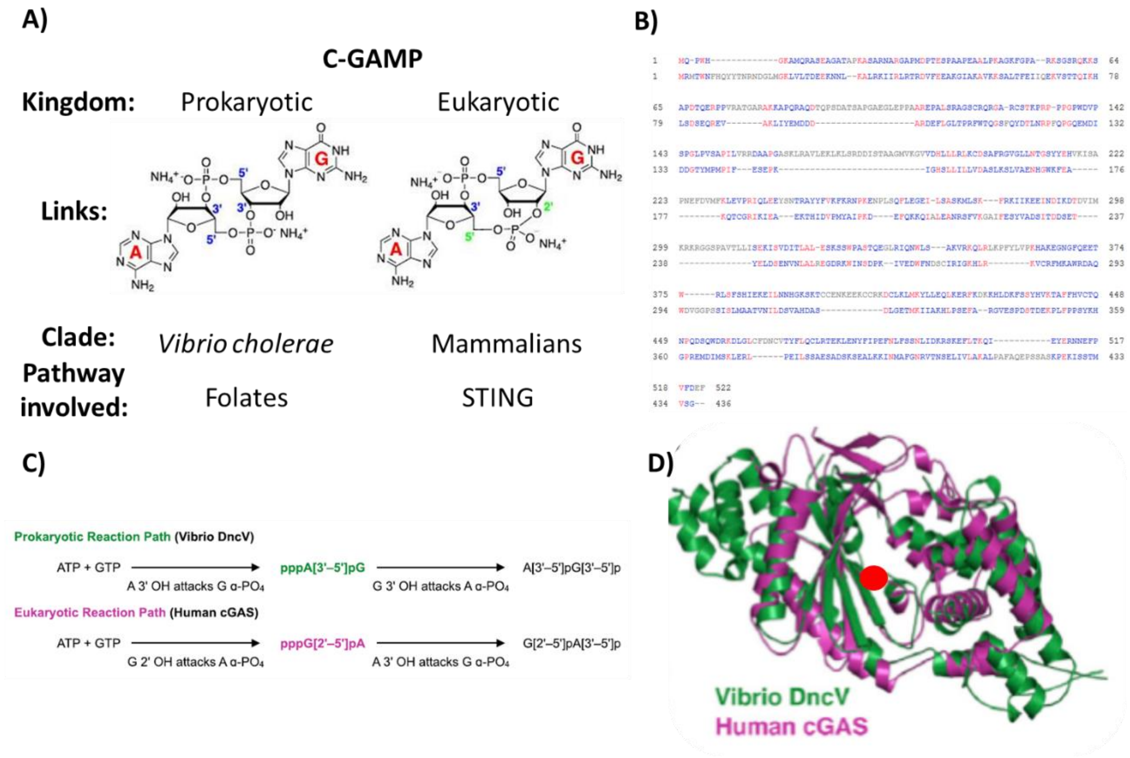


Figure 3: The recently discovered c-GAMP in its different kingdoms. A) Comparison of the prokaryotic and eukaryotic c-GAMP, B) Alignment of DncV and cGAS showing the low identity between the two proteins, C) Reactions showing the opposite process to form c-GAMP by DncV and cGAS (Kranzusch et al. 2014), D) Structural superposition of the DncV and cGAS proteins showing their structural similarity, remarkably in their active sites (red point) (Kranzusch et al. 2014).

Interestingly these two proteins (cGAS and DncV) responsible for synthesizing c-GAMP share less than 20% of identity in their primary structure (Fig. 3B). However, they are structurally very similar as shown in Fig. 3D. The active site of the catalytic domain is similar in the two proteins, and both share a double-stranded DNA-binding region (on the protein surface opposite the active site), although only cGAS binds dsDNA (Kranzusch et al., 2014).

Moreover, it has been shown that the synthesis reactions of c-GAMP proceed in opposite directions in these two enzymes. In the case of *Vibrio* DncV, the reaction begins with a nucleophilic attack of 3´OH of ATP on the alpha phosphate of GTP, followed by another nucleophilic attack, this time by the GTP 3´OH on the alpha phosphate of the ATP. For the human cGAS enzyme, the first reaction is an attack of the 2´OH of GTP on the ATP alpha phosphate followed by the nucleophilic attack of the ATP 3´OH on the alpha phosphate of GTP (Fig. 3C) (Kranzusch et al., 2014).

Comparison of the crystal structures of DncV and cGAS has permitted the design of mutants that reprogram the cGAS enzyme to produce the same product as the DncV enzyme (Kranzusch et al., 2014). In that study, the R376I mutation eliminated specific interactions with the guanine base causing base rotation and ribose pseudorotation that permits the nucleophilic attack of the GTP 3´OH on the alpha phosphate of the ATP.

Apart from the DncV and cGAS proteins, GacA from *Geobacter sulfurreduscens*, a GGDEF domain-containing protein able to bind GTP or ATP in its substrate binding site, was shown to preferentially produce cGAMP when the ATP and GTP are present in the solution (Hallberg et al., 2016). In this bacterium, cGAMP binds to specific riboswitches stimulating the translation of genes involved in pilus formation and polysaccharide biosynthesis. This regulation seems to be involved in the capacity of *Geobacter* to reduce Fe(III) oxide particles, since the knock-out of *gacA* is defective in this process (Hallberg et al., 2019).

1.1.4.- C-di-GMP:

The third cyclic dinucleotide, and up to now apparently most ubiquitous, is c-di-GMP (Fig. 4). It was first described as a regulator of the cellulose biosynthesis pathway in *Gluconobacter xylinus* (Ross et al., 1987), but now is recognized to regulate a large number of phenotypes in the transition between the sessile and motile states of principally Gram-negative bacteria (Jenal & Malone, 2006). In several bacterial species, including: *Caulobacter crescentus*, *Yersinia pestis*, *Vibrio parahaemolyticus*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, *E. coli*; it was demonstrated that high concentrations of c-di-GMP promote the expression of adhesive matrix components and result in multicellular behavior and biofilm formation while low concentrations promote unicellular behaviors such as motility (Jenal & Malone, 2006). These activities are regulated by c-di-GMP binding to a variety of receptor proteins that include PilZ domains, transcription regulators of the CAP-like family (Clp), c-di-GMP riboswitches, proteins of the NtrC family, degenerate EAL domains and others (Amikam & Galperin, 2006; Hickman & Harwood, 2008; Leduc & Roberts, 2009; Nelson et al., 2013).

The regulation of the intracellular levels of c-di-GMP is temporally and spatially controlled by proteins with GGDEF domains, that produce c-di-GMP from two molecules of GTP (Schirmer, 2016), and by proteins with phosphodiesterase activity (EC: 3.1.4.52) like HD-GYP domains, that degrade the GTP into two molecules of GMP (Crossman et al., 2017), and EAL domains, that degrade c-di-GMP into the linear molecule pGpG (Bobrov et al., 2005; Christen et al., 2005).

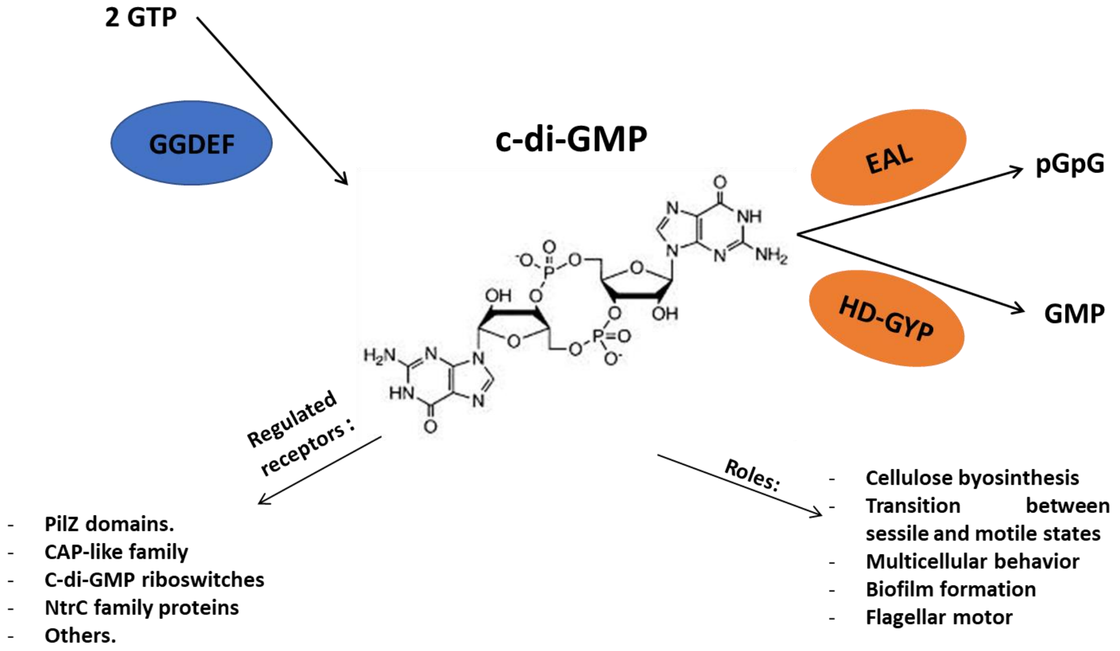


Figure 4: c-di-GMP metabolism. The c-di-GMP is produced by proteins with GGDEF domains and degraded by EAL or HD-GYP domains, the c-di-GMP achieve its several roles through the binding to its regulated receptors mention, both the roles and the regulated receptors are mention in the figure.

1.1.5.- GGDEF-containing protein, PleD:

One of the most thoroughly studied proteins with a canonical GGDEF domain is the dimeric PleD protein of *Caulobacter crescentus.* Its monomer architecture consists of two Rec/CheY domains (REC1 and REC2) and one C-terminal GGDEF domain (Fig. 5A).

The GGDEF domain consists of five beta strands surrounded by four alpha helices, a topology similar to the catalytic nucleus of the adenylate cyclase domains and the palm-like domain of the DNA polymerases (Chan et al., 2004).

PleD has a very interesting mechanism of regulation. First, to allow the catalysis and the production of c-di-GMP the protein must necessarily adopt a dimeric form, because each monomer recognizes one GTP molecule. To produce a stable dimeric form, the phosphorylation of residue D53 in the REC1 domain switches the protein to an active state by inducing a conformational change of the two REC domains promoting more contacts between the REC1 domain of one monomer and the REC2 domain of the other monomer (Fig. 5B) (Wassmann et al., 2007). It is important to note that in its inactive (non-phosphorylated) form the PleD protein can dimerize, but with much lower affinity (a 100-fold greater dissociation constant).

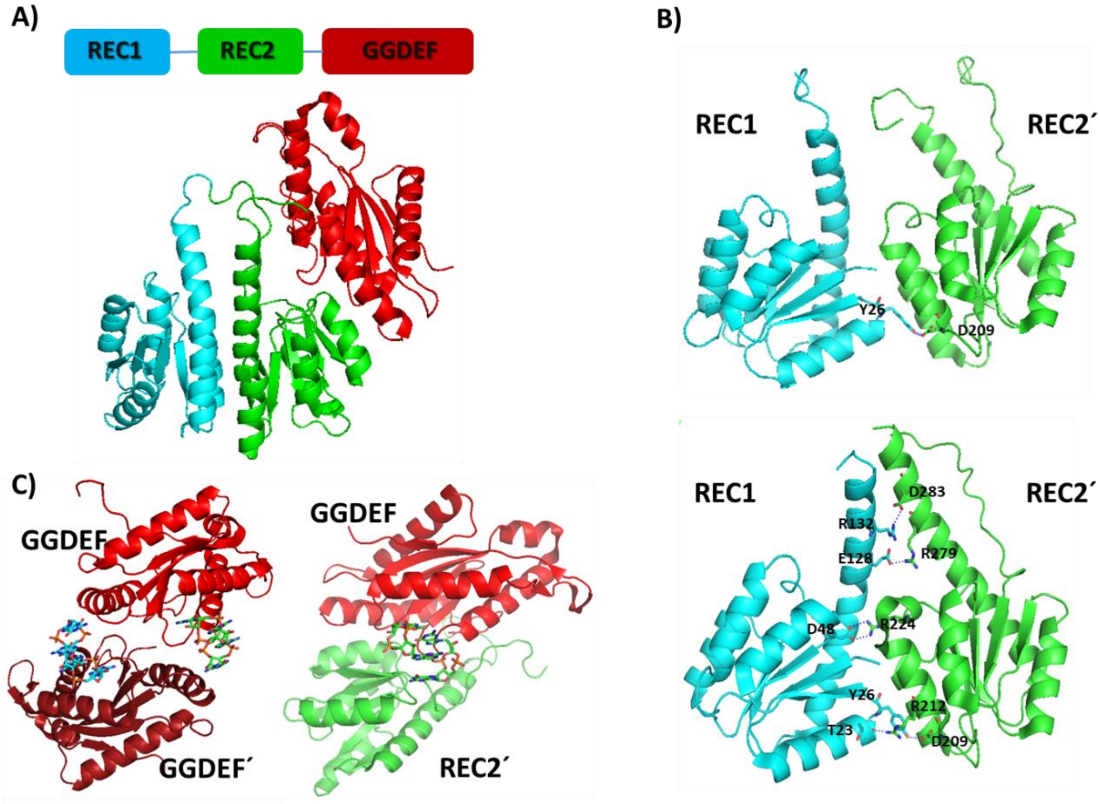


Figure 5: Structural studies in PleD (Wassmann et al. 2007 and Chan et al. 2004). A) Above, domains architecture of PleD protein; down, crystallography structure of the full protein PleD, the colors correspond to the domains in the architecture scheme. B) Comparison of the contact in a dimer of PleD between REC1 and REC2´ in the unphosphorylated state (inactive) (above) and in the phosphorylated state (active) (down). C) Allosterically inhibition by product in PleD; at left, inhibition by cross-linking of the two GGDEF domains in the dimer; and at right, inhibition by crosslinking between the REC2 and the GGDEF domain of the same monomer, both forms abolish the possibility that the two active sites can become sufficiently near.

Secondly, the PleD protein binds its product, c-di-GMP both in its phosphorylated dimeric form and in its non-phosphorylated monomeric form, in both cases çeaging to inhibition of enzymatic activity. In the non-phosphorylated form, two molecules of c-di-GMP bind to the primary inhibition site (I-site$_p$), represented by the RxxD motif in the GGDEF domain, and to the secondary inhibition site present in the REC2 domain (I-site$_{REC2}$) inducing a rigid association of the GGDEF domain with the REC2 domain that prevents the approximation of the two GGDEF domains to form a competent active site (Fig 5C right) (Chan et al., 2004). In its phosphorylated form, the c-di-GMP molecules bind to the I-site$_p$ of one of the monomers and to a second inhibitory site also present in the GGDEF domain (I-site$_{GGDEF}$), that produce a rigid association of the two GGDEF domains in an orientation that is also not catalytically active (Fig. 5C left) (Wassmann et al., 2007).

1.1.6.- GGDEF domain interaction with GTP:

With the resolution of the crystal structure of PleD (PDB ID: 2VON) in complex with its product c-di-GMP and with a non-hydrolyzable substrate (GTPαS), the residues involved in the recognition of the GTP molecule and in the coordination of the phosphate groups by magnesium ions in the active site were identified (Fig. 6).

The specificity of PleD for GTP is explained, in part, by contacts of the guanine base with the protein through two residues: N335 forms two hydrogen bonds, one with N3 and the other with the N2, whereas D344 forms another two hydrogen bonds with N2 and N1 (Wassmann et al., 2007). On the other hand, the beta and gamma phosphate groups are coordinated with the main chain amino groups of two hydrophobic residues: F330 and F331. They are also linked to a magnesium ion, which in turn is coordinated to the carbonyl group of I328 and with two highly conserved acidic GGDEF domain residues: D327 and E370 (Wassmann et al., 2007).

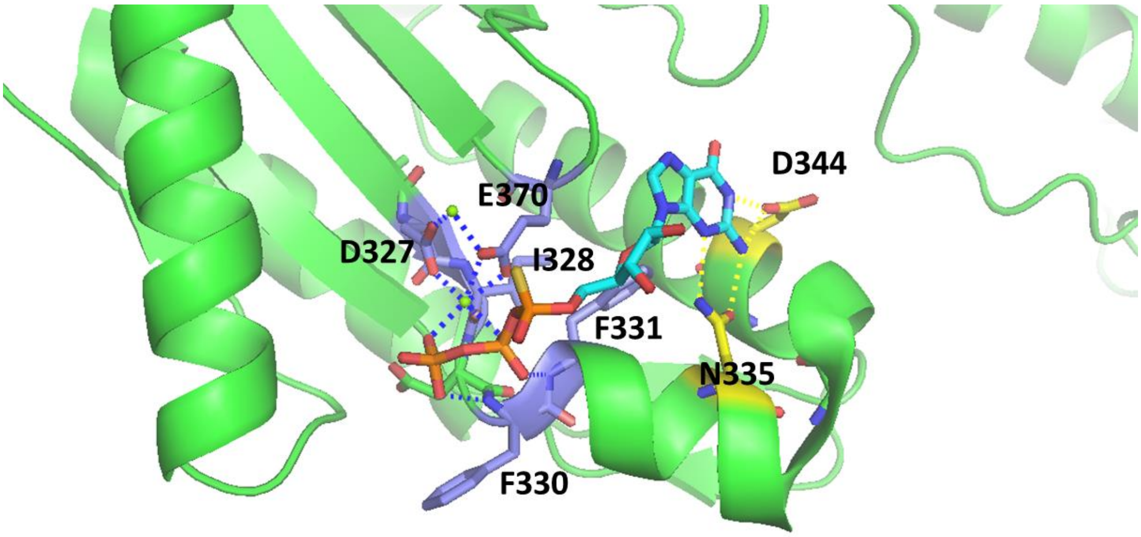


Figure 6: Analog substrate of GTP (GTPαS) bound to the active site of the GGDEF domain of PleD, there are specific contacts between the side chains of D344 and N335 with the nucleotide base. The residues F330 and F331 bind to the phosphates, and the residues D327, I328 and E370 also coordinated the phosphates through the magnesium ions.

1.1.7.- Diversity of GGDEF domains:

GGDEF domains are ubiquitous in bacteria: 90502 proteins with this domain are listed in the PFAM database (http://pfam.xfam.org/family/GGDEF) as of September 14, 2021, making this domain family one of the largest known. Furthermore, this domain is rebundant in most prokaryotes' genomes (in average each prokaryotic genome codifies 16 GGDEF domain-containing proteins, http://pfam.xfam.org/family/GGDEF). For instance, the genome of *Xanthomonas citri* 306 presents 30 GGDEF domain-containing proteins.

As mentioned above, canonical diguanylate cyclases have a GGDEF domain, with an active site called A-site (represented in part by the GGDEF motif), a feedback inhibition site (I-site, represented by the RxxD motif) and other domains associated with DGC regulation or downstream enzymatic activity. However, many GGDEF domains in the database are most likely inactive – the so-called degenerate GGDEF domains due to mutations in their A-sites. Also, many GGDEF domains lack an I-site or adjacent domains with clear regulatory functions. Some of these proteins can retain the ability to bind c-di-GMP or GTP and activate adjacent domains by inducing conformational (allosteric) changes (Jenal & Malone, 2006; Whitney et al., 2012). Finally, some GGDEF domains that have both a degenerate I-site unable to bind GTP or c-di-GMP and degenerate A-site unable to bind GTP, may have been preserved after successive mutations across generations due to a structural role that maintains other domains in the protein active. However, it is curious that a few of these degenerate domains have been shown to have new unexpected enzymatic activities: for example, the YybT protein of *Bacillus subtilis* has the capacity to recognize and degrade ATP (Rao et al., 2010), the GGDEF domain of Slr1143 in *Synechocystis sp.* can degrade GTP to GDP (Ryjenkov et al., 2005), and, as mentioned above, GSU1658 (GacA) from *Geobacter sulfurreduscens* is able to recognize GTP and ATP in different monomers to form cGAMP (Hallberg et al., 2019).

1.1.7.- Working hypothesis:

In summary, the presence of GGDEF domains with unknown function, the condition that the GGDEF domains function as a dimers, the redundance of this domain in the prokaryotic genomes, and the existence of GGDEF proteins with similar domain architectures in the same genome, lead us to hypothesize that naturally occurring dimerization between two proteins with GGDEF domains (more likely if have similar architectures), one with an active GGDEF domain and other with an inactive GGDEF domain could possibly produce a pppGpG molecule (Fig. 7B), which may exercise a signaling role. Furthermore, the formation of a dimer containing two GGDEF domains able to recognize ATP the dimer would produce c-di-AMP (Fig. 7C). Finally, the association of two active GGDEF domains, one able to bind ATP and another one able to bind GTP, would produce the hybrid molecule c-GAMP (Fig. 7D).

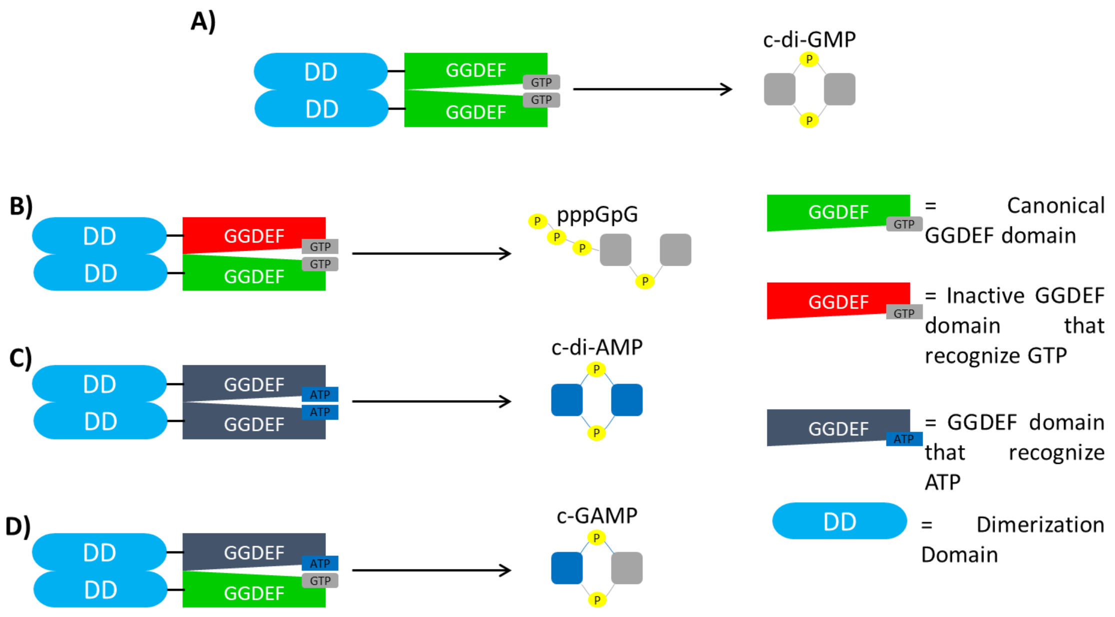


Figure 7: Hypothesis about new different enzymatic functions of the GGDEF domains. A) Canonical enzymatic activity of GGDEF domain. B) Production of a pppGpG molecule for a GGDEF dimer with one inactive monomer and one active monomer. C) Production of c-di-AMP by GGDEF domains that recognize ATP molecule. D) Production of the hybrid c-GAMP by a GGDEF dimer that can recognize ATP with one monomer and GTP with another monomer.

1.2.- OBJECTIVES:

The general objective of this project was to explore whether GGDEF domains are able to form products other than the canonical c-di-GMP. To achieve this, we put forward the following specific objectives:

- Determine the structural effects of point mutations on the active site of GGDEF domains designed to change the specificity for GTP to ATP.
- Determine the enzymatic specificity of wild-type and mutants of different DGCs; specifically exploring their ability to employ as substrate not only GTP, but also ATP, ITP, dGTP, dATP, and dITP.
- Develop a kinetic model to explain the production of non-canonical products of GGDEF domains; specifically, the formation of pppGpdG by XAC0610 from *X. citri*.

1.3.- MATERIAL AND METHODS:

1.3.1. Cloning of expression plasmids

The coding sequence for PleD* protein was obtained from the pRP89 plasmid (Paul et al., 2004) kindly donated by the Urs Jenal group at the Biozentrum – University of Basel. This plasmid has the PleD* version of the protein with a C-terminal 6x-His-tag in a pET11 vector.

To obtain an expression vector of PleD we first used genomic DNA from *Caulobacter crescentus* (kindly donated by Prof. Aline Maria da Silva) to amplify the coding region of PleD and then it was cloned in pET28a(+) (Figure 8) using the F_PleDwt28_NdeI (5´

gggcatatgAGCGCCCGGATCCTC 3´) and R_PleDwt28_XhoI primers (5´ aaactcgagGGCGGCCTTGCCGAC 3´). Finally PleD was amplified from this vector and cloned in pRP89 using the primers F_PleDwt28_NdeI and R_PleDwt11_EcoRI (5´ tttgaattcAGTGGTGGTGGTGGTGGTG 3´), obtaining a c-terminal 6x-His tagged PleD.

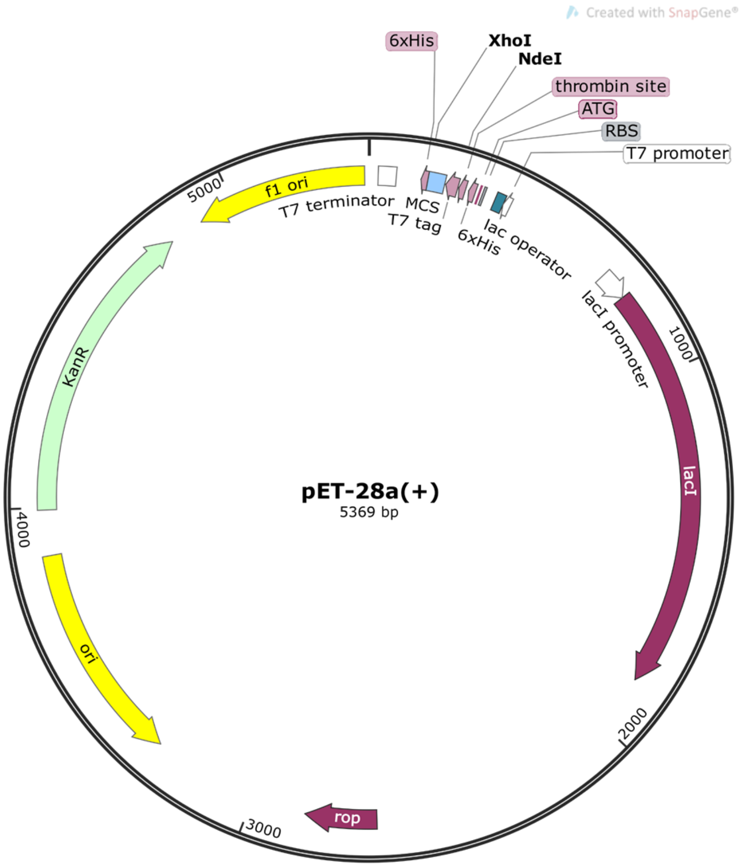


Figure 8: pET-28 plasmid map showing the relevant restriction sites and features.

XAC0610 constructions were previously created by our group (Oliveira et al., 2015). XAC0610$_{35-880}$ and XAC0610$_{701-880}$ were cloned using F_XAC0610$_{35-880}$_NdeI (5´ gaattaatcatatgGACGACGCCTTGCGCG 3´) and R_XAC0610$_{35-880}$_HinDIII (5´ gtttaagcttCTACGATCGAGGCGCG 3´) or F_XAC0610$_{701-880}$_NdeI (5´ gaactaccatatgGATGTCACTGCGCACAAGAC 3´) and R_XAC0610$_{701-880}$_HinDIII (5´

gtttaagcttCTACGATCGAGGCGCG 3´) primers respectively in the pET28a(+) vector with an N-terminal 6x-His tag. XAC0424 and XAC2810 coding sequences were amplified from *X. citri* 306 genomic DNA by PCR using the primers F_XAC0424_NcoI (5´ gtcatgCCATGGcagaccagcccgaac 3´), R_XAC0424_XhoI (5´ tataagcttCTCGAGtattaatcggcgttgatcacccgg 3´), F_XAC2810_NcoI (5´ gatttCCATGGatggtggattcgccggccg 3´), and R_XAC2810_XhoI (5´ gtccCTCGAGtactccagcgcaaccacc 3´) and cloned in the pROEX-b (Figure 9) vector using the NcoI and XhoI restriction sites, also obtaining N-terminal 6x-His tagged proteins.

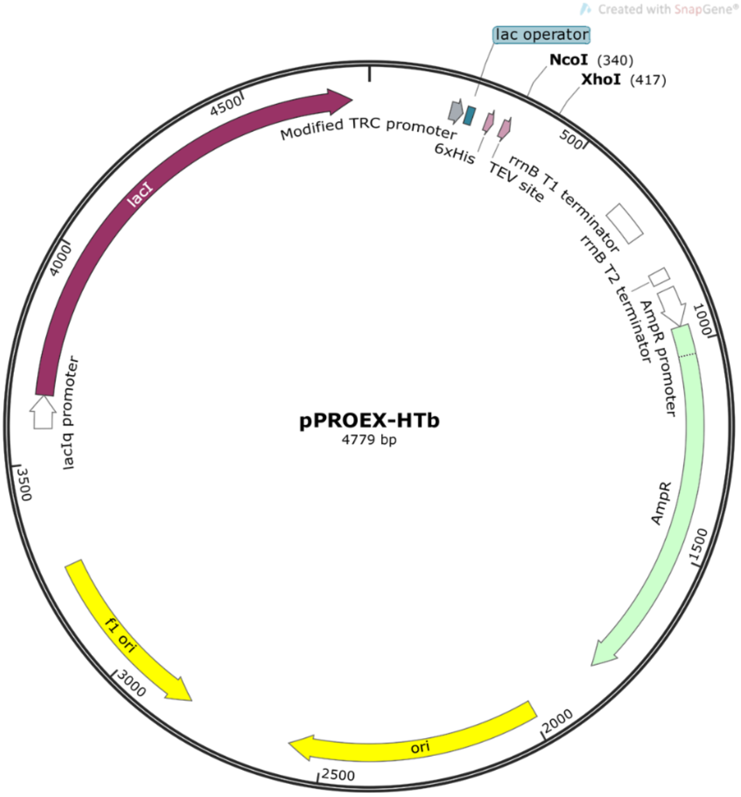


Figure 9: pPROEX-HTb plasmid map showing the relevant restriction sites and features.

Plasmids containing the full-length coding sequence of LIC11128 and LIC11131 with a N-terminal 6x-His-tag in the pOPINF vector were kindly donated by Profa. Dra. Cristianne Guzzo from the Instituto de Ciencias Biomedicas – USP.

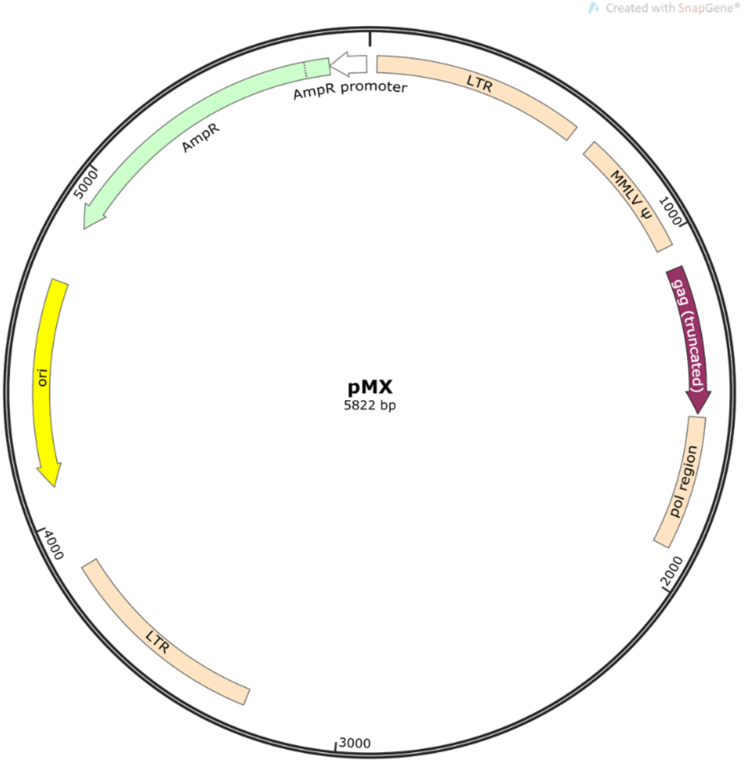


Figure 10: pMX plasmid map showing the relevant features, no restriction sites are showing here due to sequences inserted in this plasmid were directly synthetized and not cloned by restriction sites.

The coding sequence for c-terminal 6x-His GSU1658 was purchased from Invitrogen inserted in a pMX (Figure 10) plasmid. Then it was cloned in the NdeI and EcoRI sites of pRP89 plasmid (Paul et al., 2004).

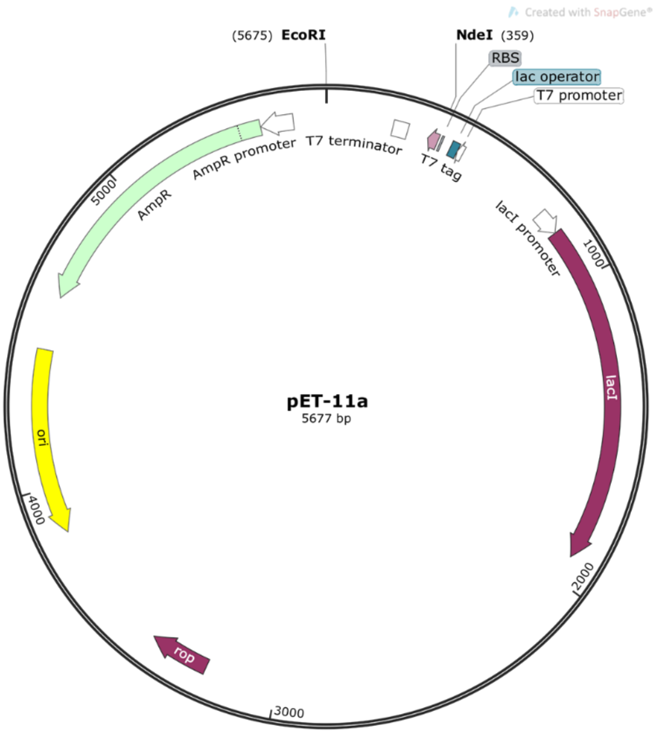


Figure 11: pET-11 plasmid map showing the relevant restriction sites and features.

The construction of PleD-GSU chimera was done by Gibson Assembly reactions using three PCR products: i) Using a pET28 plasmid containing the coding sequence of PleD between the NdeI and XhoI restriction sites as a template we amplified two products using the following primers: F1_Vec$_{\text{PleD28GSU\_GGDEF(297-458)}}$ (5´ gatccggctgctaacatggc 3´), R1_Vec$_{\text{PleD28GSU\_GGDEF(297-458)}}$ (5´ GAATGGTCTTCGGTTTCCGTGTTTCGTAAAgtctggaaacgcggaagtca 3´) and F2_Vec$_{\text{PleD28GSU\_GGDEF(297-458)}}$ (5´ tttacgaaacacggaaaccg 3´), R2_Vec$_{\text{PleD28GSU\_GGDEF(297-458)}}$ (5´ CAGATAGCGATAATTGAACAGACCTGTCAGctggtcggtgacggccagct 3´), ii) Using a pET11 (Figure 11) plasmid containing the GSU1658 coding sequence cloned in the NdeI and EcoRI sites as a template we amplified a product using the following primers: F_GSU_GGDEF(297-458)_PleD_28 (5´ ctgacaggtctgttcaatta 3´), R_GSU_GGDEF(297-458)_PleD_28 (5´ CGTCCCATTCGCCATGTTAGCAGCCGGATCtcaatgatgatggtgatgat 3´).

This Gibson assembly reaction produced a pET28 plasmid containing a coding sequence of aminoacids 1 – 293 of PleD and aminoacids 297 – 458 of GSU1658 with a C-terminal 6xHis-tag. Finally, using the primers F_wt_PleD (5´ gggcatATGAGCGCCCGGATCCTC 3´) and R_PleD_GSU (5´ tttgaatTCAATGATGATGGTGATGATGACG 3´) we amplified this region and cloned it in a pET11 plasmid in NdeI and EcoRI restriction sites. The sequence of the coding region was verified by sequencing.

1.3.2. Expression and purification of recombinant proteins:

To express PleD* protein, we used the *E. coli* BL21(DE3) pLysS strain, induced by 0.3 mmol/L of IPTG for 23 hours at 18°C. The cells were lysed by sonication using a lysis buffer (100 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 15 mmol/L $MgCl_2$, 50 mmol/L imidazole) and the resulting suspension was clarified by centrifugation (13000 x g). The supernatant fluid was loaded onto Ni-NTA affinity resin (His-trap chelating 5 mL column), washed with 10 volumes of lysis buffer, and eluted with elution buffer (100 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 15 mmol/L $MgCl_2$, 250 mmol/L Imidazole). The eluted fraction was further purified by size exclusion chromatography using a Superdex pg 200 16/600 column and 100 mmol/L Tris-Cl pH 8.00, 100 mmol/L NaCl and 15 mmol/L $MgCl_2$ as running buffer. The eluted fractions were examined for purity by SDS-PAGE and fractions containing pure protein were concentrated in Amicon Ultra-15 Centrifugal filter units. To determine the protein concentration, we use the BCA and Hartree Lowry methods.

PleD was expressed using *E. coli* BL21(DE3), induced by 0.3 mmol/L of IPTG for 18 hours at 18°C. The cells were lysed by sonication using a lysis buffer (20 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 5 mmol/L $MgCl_2$, 50 mmol/L Imidazole) and the resulting suspension was clarified by centrifugation (20000 x g x 45 min). The supernatant fluid was loaded onto Ni-NTA affinity resin column (His-trap chelating 5 mL column) equilibrated with lysis buffer, washed with 20 volumes of lysis buffer and eluted with elution buffer (20 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 5 mmol/L $MgCl_2$, 250 mmol/L Imidazole). The eluted fraction was further purified by size exclusion chromatography using a Superdex pg 200 16/600 and 20 mmol/L Tris-Cl pH 8.00, 100 mmol/L NaCl and 5 mmol/L $MgCl_2$ as running buffer. The eluted fractions were examined for purity by SDS-PAGE and fractions containing pure protein were concentrated in Amicon Ultra-15 Centrifugal filter units. To determine the protein concentration, we used the BCA method.

Other proteins used in this work were expressed using *E. coli* BL21(DE3), induced by 0.3 mmol/L of IPTG for 18 hours at 18°C. The cells were lysed by sonication using a lysis buffer (20 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 5 mmol/L $MgCl_2$, 20 mmol/L Imidazole) and the resulting suspension was clarified by centrifugation (20000 x g x 45 min). The supernatant was loaded onto Ni-NTA affinity resin column (His-trap chelating 5 mL column) equilibrated with lysis buffer, washed with 30 volumes of lysis buffer, and eluted in a 20 – 500 mmol/L Imidazole gradient (20 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl, 5 mmol/L $MgCl_2$). Fractions containing the protein of interest were concentrated using Amicon Ultra-15 Centrifugal filter up to 5 mL and further purified by size exclusion chromatography using a Superdex pg 75 26/600 or Superdex pg 200 26/600 columns and 20 mmol/L Tris-Cl pH 8.00, 100 mmol/L NaCl and 5 mmol/L $MgCl_2$ as running buffer. The eluted fractions were examined for purity by SDS-PAGE and fractions containing pure protein were concentrated in Amicon Ultra-15 Centrifugal filter units. Protein concentration was determined using UV absorbance at 280 nm.

1.3.3. High-throughput cloning, expression, and purification of diguanylate cyclase XAC0610 constructions:

Primers were designed to amplify and clone in pOPINF vectors (Figure 12) all the possible combinations of XAC0610 domains (Table 1) and used for PCR reactions using pET28-XAC0610$_{35\text{-}880}$ as a template. First PCR reactions were done using Phusion enzyme and annealing temperature of 60 ºC. Amplicons were evaluated in 1.6 % agarose gels running in 1X TBE buffer. PCR reactions that did not show amplicons with the expected size were repeated lowering the annealing temperature to 55 ºC and the products were again evaluated in 1.6 % agarose gels running in 1X TBE buffer. PCR reactions that showed nonspecific bands were purified using an agarose-gel extraction method (GeneJET gel extraction kit, Thermo scientific). PCR reactions without non-specific bands were purified using AMPure XP magnetic bead purification method. PCR amplicons were cloned using the Bioquote cloning kit. To perform this, we mix 0.5 ul of 100 ng/ul linearized pOPINF vector, 1 ul of PCR product, 1 ul of 5x buffer, 0.5 ul of Quick-fusion enzyme and 2 ul of water. These reactions were incubated at 37 ºC for 30 minutes. 2 ul of these reactions were used to transform Stellar ultracompetent *E. coli* cells. Positive colonies were used to isolate plasmids and we confirm the presence of the insert in the plasmid by PCR. Those plasmids were used to transform *E. coli* BL21(DE3) LEMO21 cells for subsequent expression assay. For expression assays, individual colonies were used to inoculate power broth medium supplemented with chloramphenicol and ampicillin and incubated overnight at 37 ºC. After that, 150 ul of the overnight cultures were used to inoculate 3 mL of Overnight Express^{TM} Instant TB medium (TBONEX) supplemented with ampicillin and chloramphenicol. These cultures were incubated at 37 ºC at 200 rpm for approximately 4 hours and then the incubation temperature was reduced to 25 ºC degrees. This temperature was further maintained for 20 more hours. After that, 1 mL of each culture was transferred to 96-well deep-well blocks and centrifuged at 6000 x g for 10 minutes. The supernatant was discarded, and the cells were stored at -80 ºC for 30 minutes. Then, the cells were defrosted and resuspended in 210 ul of Lysis buffer (50 mmol/L NaH2PO4, 300 mmol/L NaCl, 10 mmol/L Imidazole, 1 % v/v Tween 20, pH 8.0) supplemented with 1 mg/mL lysozyme and 3 U/mL of Benzonase. After 30 minutes in room temperature the well block was centrifuged at 6000 g for 30 minutes at 4 ºC. The supernatant was transferred to a new well block containing 20 ul of Ni-NTA magnetic bead suspension and was mixed for 30 minutes at room temperature. The well block was then placed in the well magnet for 1 minute and the supernatant was carefully discarded. Then, the magnetic beads were washed twice with 200 ul of wash buffer (50 mmol/L NaH2PO4, 300 mmol/L NaCl, 20 mmol/L imidazole, 0,05 % v/v tween 20, pH 8.0). Finally, 50 ul of elution buffer (50 mmol/L NaH2PO4, 300 mmol/L NaCl, 250 mmol/L imidazole, 0,05 % v/v tween 20, pH 8.0) was added, samples were vortexed for 1 minute, placed in the magnet for 1 minute and the supernatant was transferred to clean 96-well PCR tubes. These samples were analyzed by SDS-PAGE.

Table 1. List of primers used in the high-throughput screening of 47 constructions of XAC0610 study.

Domain Coverage	aa_N	aa_C	Fwd primer
GAF + PAS domains + GGDEF	48	877	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF + PAS domains + GGDEF	46	879	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
GAF	48	188	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF	35	188	AAGTTCTGTTTCAGGGCCCGGACGACGCCTTGCGCGTG
GAF	46	190	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
PAS1	200	324	AAGTTCTGTTTCAGGGCCCGTTGAGCATGCTGCTGGAAGC
PAS1	198	326	AAGTTCTGTTTCAGGGCCCGCAGACCTTGAGCATGCTGCTG
PAS2	332	461	AAGTTCTGTTTCAGGGCCCGCTGCAGGCACTGGTGGATG
PAS2	330	463	AAGTTCTGTTTCAGGGCCCGGCGCGGCTGCAGGCACTG
PAS3	472	580	AAGTTCTGTTTCAGGGCCCGGCATTCGAAACCGCGCCGC
PAS3	470	582	AAGTTCTGTTTCAGGGCCCGGCCGGTGCATTCGAAACCG
PAS4	596	711	AAGTTCTGTTTCAGGGCCCGCTGCGCGCGATCAGCGAC
PAS4	594	713	AAGTTCTGTTTCAGGGCCCGGCGCGCCTGCGCGCGATC
GGDEF	712	877	AAGTTCTGTTTCAGGGCCCGATGCACGAGCGCGCCACC
GGDEF	710	879	AAGTTCTGTTTCAGGGCCCGCGCCTGATGCACGAGCGC
GAF+PAS1	35	324	AAGTTCTGTTTCAGGGCCCGGACGACGCCTTGCGCGTG
GAF+PAS1	48	324	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF+PAS2	46	326	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
PAS1+PAS2	200	461	AAGTTCTGTTTCAGGGCCCGTTGAGCATGCTGCTGGAAGC
PAS1+PAS2	198	463	AAGTTCTGTTTCAGGGCCCGCAGACCTTGAGCATGCTGCTG
PAS2+PAS3	332	580	AAGTTCTGTTTCAGGGCCCGCTGCAGGCACTGGTGGATG
PAS2+PAS3	330	582	AAGTTCTGTTTCAGGGCCCGGCGCGGCTGCAGGCACTG
PAS3+PAS4	472	711	AAGTTCTGTTTCAGGGCCCGGCATTCGAAACCGCGCCGC
PAS3+PAS4	470	713	AAGTTCTGTTTCAGGGCCCGGCCGGTGCATTCGAAACCG
PAS4+GGDEF	596	877	AAGTTCTGTTTCAGGGCCCGCTGCGCGCGATCAGCGAC
PAS4+GGDEF	594	879	AAGTTCTGTTTCAGGGCCCGGCGCGCCTGCGCGCGATC
GAF+PAS1+PAS2	35	461	AAGTTCTGTTTCAGGGCCCGGACGACGCCTTGCGCGTG
GAF+PAS1+PAS2	48	461	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF+PAS1+PAS2	46	463	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
PAS1+PAS2+PAS3	200	580	AAGTTCTGTTTCAGGGCCCGTTGAGCATGCTGCTGGAAGC
PAS1+PAS2+PAS3	198	582	AAGTTCTGTTTCAGGGCCCGCAGACCTTGAGCATGCTGCTG
PAS2+PAS3+PAS4	332	711	AAGTTCTGTTTCAGGGCCCGCTGCAGGCACTGGTGGATG
PAS2+PAS3+PAS4	330	713	AAGTTCTGTTTCAGGGCCCGGCGCGGCTGCAGGCACTG
PAS3+PAS4+GGDEF	472	877	AAGTTCTGTTTCAGGGCCCGGCATTCGAAACCGCGCCGC
PAS3+PAS4+GGDEF	470	879	AAGTTCTGTTTCAGGGCCCGGCCGGTGCATTCGAAACCG
GAF+PAS1+PAS2+PAS3	35	580	AAGTTCTGTTTCAGGGCCCGGACGACGCCTTGCGCGTG
GAF+PAS1+PAS2+PAS3	48	580	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF+PAS1+PAS2+PAS3	46	582	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
PAS1+PAS2+PAS3+PAS4	200	711	AAGTTCTGTTTCAGGGCCCGTTGAGCATGCTGCTGGAAGC
PAS1+PAS2+PAS3+PAS4	198	713	AAGTTCTGTTTCAGGGCCCGCAGACCTTGAGCATGCTGCTG

PAS2+PAS3+PAS4+GGDEF	332	877	AAGTTCTGTTTCAGGGCCCGCTGCAGGCACTGGTGGATG
PAS2+PAS3+PAS4+GGDEF	330	879	AAGTTCTGTTTCAGGGCCCGGCGCGGCTGCAGGCACTG
GAF+PAS1+PAS2+PAS3	35	711	AAGTTCTGTTTCAGGGCCCGGACGACGCCTTGCGCGTG
GAF+PAS1+PAS2+PAS3	48	711	AAGTTCTGTTTCAGGGCCCGGTGCTGGACACCGAGGCC
GAF+PAS1+PAS2+PAS3	46	713	AAGTTCTGTTTCAGGGCCCGCTGGGGGTGCTGGACACC
PAS1+PAS2+PAS3+PAS4+GGDEF	200	877	AAGTTCTGTTTCAGGGCCCGTTGAGCATGCTGCTGGAAGC
PAS1+PAS2+PAS3+PAS4+GGDEF	198	879	AAGTTCTGTTTCAGGGCCCGCAGACCTTGAGCATGCTGCTG
Domain Coverage	**aa_N**	**aa_C**	**Rev primer**
GAF + PAS domains + GGDEF	48	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTG
GAF + PAS domains + GGDEF	46	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATC
GAF	48	188	ATGGTCTAGAAAGCTTTAGCGGCGCGCTTCCAACTTG
GAF	35	188	ATGGTCTAGAAAGCTTTAGCGGCGCGCTTCCAACTTG
GAF	46	190	ATGGTCTAGAAAGCTTTAGCGATCGCGGCGCGCTTC
PAS1	200	324	ATGGTCTAGAAAGCTTTATGCTGCGGAGGCATCCTTG
PAS1	198	326	ATGGTCTAGAAAGCTTTAGGCCAGTGCTGCGGAGGC
PAS2	332	461	ATGGTCTAGAAAGCTTTAAGCGGCCTGCGCCTGCAATTC
PAS2	330	463	ATGGTCTAGAAAGCTTTATTGCGCAGCGGCCTGCGC
PAS3	472	580	ATGGTCTAGAAAGCTTTAGGTAACGTCCTGGATCTGCG
PAS3	470	582	ATGGTCTAGAAAGCTTTAACGCTCGGTAACGTCCTGG
PAS4	596	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
PAS4	594	713	ATGGTCTAGAAAGCTTTAGTGCATCAGGCGGTGCAGG
GGDEF	712	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTG
GGDEF	710	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATC
GAF+PAS1	35	324	ATGGTCTAGAAAGCTTTATGCTGCGGAGGCATCCTTG
GAF+PAS1	48	324	ATGGTCTAGAAAGCTTTATGCTGCGGAGGCATCCTTG
GAF+PAS2	46	326	ATGGTCTAGAAAGCTTTAGGCCAGTGCTGCGGAGGC
PAS1+PAS2	200	461	ATGGTCTAGAAAGCTTTAAGCGGCCTGCGCCTGCAATTC
PAS1+PAS2	198	463	ATGGTCTAGAAAGCTTTATTGCGCAGCGGCCTGCGCC
PAS2+PAS3	332	580	ATGGTCTAGAAAGCTTTAGGTAACGTCCTGGATCTGCG
PAS2+PAS3	330	582	ATGGTCTAGAAAGCTTTAACGCTCGGTAACGTCCTGG
PAS3+PAS4	472	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
PAS3+PAS4	470	713	ATGGTCTAGAAAGCTTTAGTGCATCAGGCGGTGCAGG
PAS4+GGDEF	596	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTG
PAS4+GGDEF	594	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATC
GAF+PAS1+PAS2	35	461	ATGGTCTAGAAAGCTTTAAGCGGCCTGCGCCTGCAATTC
GAF+PAS1+PAS2	48	461	ATGGTCTAGAAAGCTTTAAGCGGCCTGCGCCTGCAATTC
GAF+PAS1+PAS2	46	463	ATGGTCTAGAAAGCTTTATTGCGCAGCGGCCTGCGC
PAS1+PAS2+PAS3	200	580	ATGGTCTAGAAAGCTTTAGGTAACGTCCTGGATCTGCG
PAS1+PAS2+PAS3	198	582	ATGGTCTAGAAAGCTTTAACGCTCGGTAACGTCCTGG
PAS2+PAS3+PAS4	332	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
PAS2+PAS3+PAS4	330	713	ATGGTCTAGAAAGCTTTAGTGCATCAGGCGGTGCAGG
PAS3+PAS4+GGDEF	472	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTGC
PAS3+PAS4+GGDEF	470	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATC
GAF+PAS1+PAS2+PAS3	35	580	ATGGTCTAGAAAGCTTTAGGTAACGTCCTGGATCTGCG
GAF+PAS1+PAS2+PAS3	48	580	ATGGTCTAGAAAGCTTTAGGTAACGTCCTGGATCTGCG
GAF+PAS1+PAS2+PAS3	46	582	ATGGTCTAGAAAGCTTTAACGCTCGGTAACGTCCTGG

PAS1+PAS2+PAS3+PAS4	200	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
PAS1+PAS2+PAS3+PAS4	198	713	ATGGTCTAGAAAGCTTTAGTGCATCAGGCGGTGCAGG
PAS2+PAS3+PAS4+GGDEF	332	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTG
PAS2+PAS3+PAS4+GGDEF	330	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATC
GAF+PAS1+PAS2+PAS3	35	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
GAF+PAS1+PAS2+PAS3	48	711	ATGGTCTAGAAAGCTTTACAGGCGGTGCAGGGTCTTG
GAF+PAS1+PAS2+PAS3	46	713	ATGGTCTAGAAAGCTTTAGTGCATCAGGCGGTGCAGG
PAS1+PAS2+PAS3+PAS4+GGDEF	200	877	ATGGTCTAGAAAGCTTTACGCGTGCCCATCTGGCTG
PAS1+PAS2+PAS3+PAS4+GGDEF	198	879	ATGGTCTAGAAAGCTTTATCGAGGCGCGTGCCCATCTG

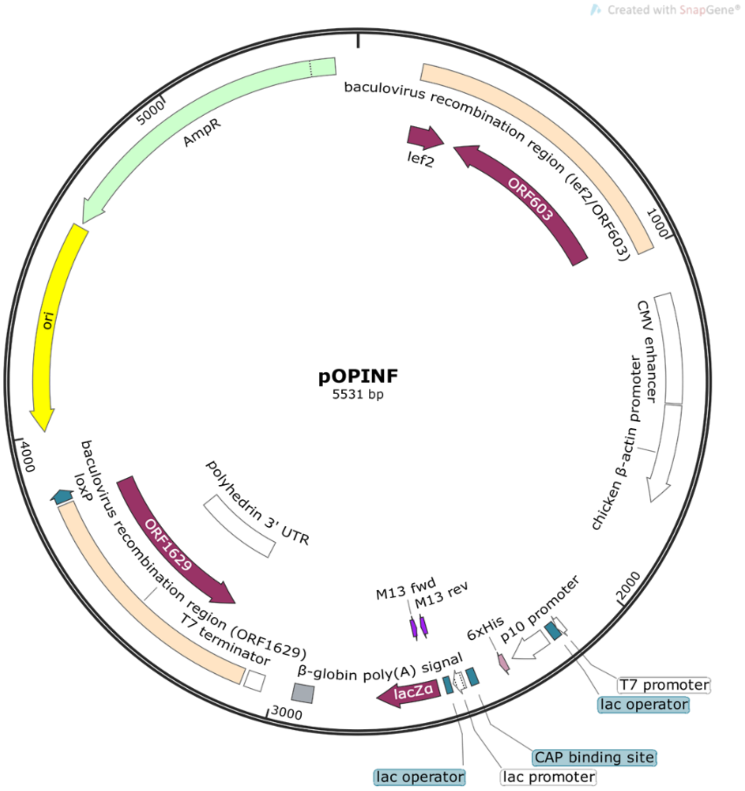


Figure 12: pOPINF plasmid map showing the relevant features, no restriction sites are showing here due to sequences inserted in this plasmid were done using the Bioquote cloning kit.

1.3.4. Induction assays of XAC0610 GGDEF domain constructions:

pOPINF plasmids produced in the OPPF (described in the previous section (1.3.3)) containing the following XAC0610 fragments: XAC0610$_{\text{PAS4-GGDEF(596-877)}}$, XAC0610$_{\text{PAS4-GGDEF(594-879)}}$, XAC0610$_{\text{PAS3-PAS4-GGDEF(472-877)}}$, XAC0610$_{\text{PAS3-PAS4-GGDEF(470-879)}}$, XAC0610$_{\text{PAS2-PAS3-}}$

$_{\text{PAS4-GGDEF(332-877)}}$, XAC0610$_{\text{PAS2-PAS3-PAS4-GGDEF(330-879)}}$, XAC0610$_{\text{PAS1-PAS2-PAS3-PAS4-GGDEF(198-879)}}$ were transformed in *E. coli* BL21(DE3) strain. Those strains were grown in 2 mL 2xTY broth in 24 well-plates at 37 ºC up to OD$_{600}$ 0.6 – 0.8. After that, IPTG was added to a final concentration of 0.5 mmol/L and the cultures were incubated at 18 ºC for 18 hours. A sample before and after the inductions was evaluated by SDS-PAGE and western blot using antibodies against 6xHis-tag.

1.3.5. Induction assays of other XAC0610 constructions:

pOPINF plasmids produced in the OPPF (described in the report of 2018) containing the following XAC0610 fragments: XAC0610$_{\text{PAS2(330-463)}}$, XAC0610$_{\text{PAS3(472-580)}}$, XAC0610$_{\text{PAS3(470-582)}}$, XAC0610$_{\text{PAS2(332-461)}}$, XAC0610$_{\text{GAF-PAS1(48-324)}}$, XAC0610$_{\text{GAF-PAS1(35-324)}}$ were transformed in *E. coli* BL21(DE3) strain. Those strains were grown in 2 mL 2xTY broth in 24 well-plates at 37 ºC up to OD$_{600}$ 0.6 – 0.8. After that, IPTG was added to a final concentration of 0.5 mmol/L and the cultures were incubated at 18 ºC for 18 hours. A sample before and after the inductions was evaluated by SDS-PAGE and western blot using antibodies against 6xHis-tag.

1.3.6 Cloning of RNA-based biosensors:

pMX (Figure 10) plasmids containing sequences of GSU1658-His, cdiAMP biosensor and cGAMP biosensor were purchased from Invitrogen. Biosensor sequences consist of a T7 promoter followed by a 5´ section of tRNA scaffold, 5´ section of spinach aptamer, the specific cyclic dinucleotide aptamer (P1-4delA Gm0970 for cGAMP and YuaA P1-4 for cdiAMP (Kellenberger, Chen, et al., 2015; Kellenberger, Wilson, et al., 2015), the 3´ section of spinach aptamer, the 3´ section of tRNA scaffold and a T7 terminator sequence flanked by a 5´ SphI restriction site and a 3´ BglII restriction site. The pMX-cGAMP biosensor plasmid was digested with SphI and BglII and the fragment corresponding to the biosensor was cloned in pET28a(+) linearized with the same enzymes. This construction was named pBiocGAMP. The entire ORF of GSU1658-His was amplified from pMX-GSU1658 plasmid using primers containing restriction sites for NdeI (in the forward primer) and EcoRI (in the reverse primer) (Table 2) and subcloned in the corresponding NdeI and EcoRI sites of pBiocGAMP plasmid, this plasmid was named pBiocGAMP-GSU1658. Similarly, pMX-cdiAMP was digested with SphI and BglII to obtain the cdiAMP biosensor sequence and was cloned in the corresponding restriction sites in pET28a(+) obtaining the pBiocdiAMP plasmid. The ORF of DisA (DNA integrity scanning protein), a diadenylate cyclase was amplified from genomic DNA of *Bacillus subtilis* strain PY79 kindly donated by Prof. Dr. Frederico José Gueiros Filho using the primers F_DisA (containing a NdeI restriction site) and R_DisA (containing a XhoI restriction site) (Table 1). This fragment was cloned in the corresponding NdeI and XhoI sites of pBiocdiAMP plasmid obtaining the pBiocdiAMP-DisA plasmid. To obtain the cdiGMP biosensor sequence we used overlapping PCR reactions of three sequences purchased as primers (Table 2) each one containing approximately one third of the cdiGMP biosensor sequence (VC2 aptamer for cdiGMP (Kellenberger et al., 2013)). PCR reactions were performed first by mixing 1_BiocdiGMP with 2_BiocdiGMP and 3_BiocdiGMP with R_BiocdiGMP. Products of these reactions were purified from agarose gel after verification of the correct size and a mixture of the two products was used as a sample for another PCR. After this second PCR reaction and agarose gel electrophoresis purification, we treated this product and the pBiocGAMP-GSU1658 plasmid with SphI and SpeI restriction enzymes and ligated them together to obtain pBiocdiGMP-GSU1658 plasmid. Finally, we digested the pRP89 plasmid (Paul et al., 2004) with NdeI and EcoRI restriction enzymes to obtain the ORF sequence of PleD* and this fragment was cloned in the respective position in pBiocdiGMP-GSU1658 obtaining pBiocdiGMP-PleD* plasmid. Sequences of BiocdiGMP (in pBiocdiGMP-PleD*) and DisA (in pBiocdiAMP-DisA) were confirmed by sequencing. Figure 13 shows the map of the pBiocdiGMP-PleD*, the plasmids maps of pBiocdiAMP-DisA and pBiocGAMP-GSU1658 are identical to pBiocdiGMP-PleD* but with different proteins or biosensor sequences.

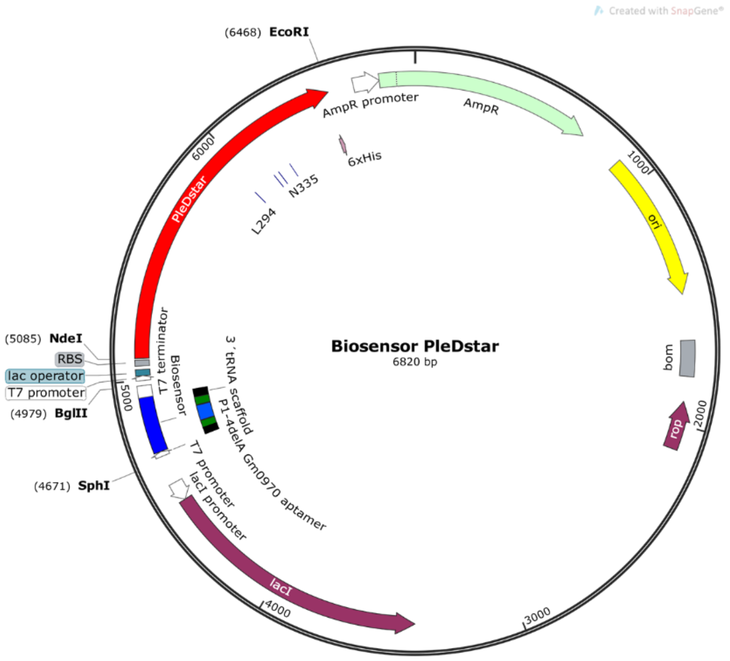


Figure 13: pBiocdiGMP-PleD* plasmid map showing the relevant restriction sites and features. The backbone is the same for pBiocdiAMP-DisA and pBiocGAMP-GSU1658 but protein sequence (in red) and biosensor sequence (in blue) are diferente according to the protein and the biosensor of interest.

Table 2. List of primers used in the biosensor constructions

Name	Sequence (5´ - 3´)
F_GSU1658	ATAcatatgGAACGTATTCTGGTGG
R_GSU1658	CAAgaattcAATGATGATGGTGATGATGACG
F_DisA	CCCcatatgGAAAAAGAGAAAAAAGGGGC
R_DisA	TTTctcgagCAGTTGTCTGTCTAAATAATGC
1_BiocdiGMP	TTTgcatgcCGATCCCGCGAAATTAATACGACTCACTATAGGGGCCCGGATAGCTCAGTCGGTAGAGCAGCGGCCGGATGTAACTGAATGAAA
2_BiocdiGMP	ttaggccggaggctttgcgtcccactctttcgaatggtttgccctgtgcgtgTGGACCCGTCCTTCACCATTTCATTCAGTTACATCCGGC
3_BiocdiGMP	gacgcaaagcctccggcctaaaccagaagacatggtaggtagcggggttaccgatgTTGTTGAGTAGAGTGTGAGCTCCGTAACTAGTggg
F_BiocdiGMP	TTTgcatgcCGATCCCGCGAAATTAATAC
R_BiocdiGMP	CCCagatctCAAAAAACCCCTCAAGACC

1.3.7. RNA-based biosensor tests:

Biosensor plasmids pBiocGAMP-GSU1658, pBiocdiAMP-DisA and pBiocdiGMP-PleD* were transformed in *E. coli* BL21(DE3) strain. Expression assays were done with these strains at OD600 between 0.6-0.8 adding IPTG at 1 mmol/L final concentration. The induction was for one hour at 37 ºC. Fluorescence microscope assays were done with BL21(DE3) strains transformed with pBiocdiGMP and pBRA vector constructs expressing XAC2382_HG (residues 198-446) or XAC2382_HGE (residues 198-705) (pBRA plasmids construction described in Teixeira et al., 2018). For these assays, we grew the strains in 5 mL of 2XTY broth at 37 ºC until an OD600 between 0.6 - 0.8 followed by the addition of 1 mmol/L IPTG and 0.6 % arabinose and the strains were incubated for 2 hours at 37 ºC. Then, 100 ul aliquots of each culture tested were centrifuged at 10000 rpm for 2 minutes and the pellet was washed twice with 500 ul PBS pH 7.0. After that, 1 ul sample was placed in a glass slide with a fine layer of 1 % agarose and 50 µmol/L DFHBI in PBS pH 7.0. Finally, the slides were incubated for 90 minutes at 37 ºC and then phase contrast and msfGFP excitation images were obtained with a LEICA DMI-8 epi-fluorescent microscope. The microscope was equipped with a HC PL APO 100x/1.4 Oil ph3 objective (LEICA) and GFP excitation-emission band-pass filter cube (Ex?470-40, DC? 495, EM? 525/50/ LEICA)

1.3.8. Point Mutations:

Point mutations in the PleD*, PleD and XAC0610 genes were introduced using the QuickChange II Site-Directed Mutagenesis Kit (Agilent) following the manufacturer's instructions. The design of the mutants was based on the crystal structure of PleD (PDB: 2VON) and the primers used are shown in table 3.

Table 3. List of primers used in the generation of point mutations in XAC0610 and PleD.

Name	Sequence (5'-3')
PleD	
N335T_1	5'-gacatcgatttcttcaagaaaatca**c**cgacaccttcggt-3'
N335T_2	5'-accgaaggtgtcg**g**tgattttcttgaagaaatcgatgtc-3'
D344S_1	5'-cgcgcagcacctcg**ct**gccgatatcgtgac-3'
D344S_2	5'-gtcacgatatcggc**ag**cgaggtgctgcgcg-3'
D344N_1	5'-cgcgcagcacctcgt**t**gccgatatcgtgac-3'
D344N_2	5'-gtcacgatatcggc**a**acgaggtgctgcgcg-3'
L294E_1	5'-gtgcaggccggtctcctggtcggtgacg-3'
L294E_2	5'-cgtcaccgaccaggagaccggcctgcac-3'
L294Q_1	5'-gtgcaggccggtctgctggtcggtgacg-3'
L294Q_2	5'-cgtcaccgaccagcagaccggcctgcac-3'

N14Y_1	5'-agcaggcggacataggcctcgatgtcg-3'
N14Y_2	5'-cgacatcgaggcctatgtccgcctgct-3'
XAC0610	
D771R_1	5'-GGTCACCGTGCCGGCcgTGCGGTGCTGGTGGCG-3'
D771R_2	5'-CGCCACCAGCACCGCAcgGCCGGCACGGTGACC-3'
D771S_1	5'-caccagcaccgcACTgccggcacggtga-3'
D771S_2	5'-tcaccgtgccggcAGTgcggtgctggtg-3'
L721Q_1	5'-cggcaggccggtCTGggcatcgcgggtg-3'
L721Q_2	5'-cacccgcgatgccCAGaccggcctgccg-3'
L721E_1	5'-cggcaggccggtCTCggcatcgcgggtg-3'
L721E_2	5'-cacccgcgatgccGAGaccggcctgccg-3'
R793H_1	5'-GTTATCTGGTGGCGCaCCTGGCCGGCGACG-3'
R793H_2	5'-CGTCGCCGGCCAGGtGCGCCACCAGATAAC-3'

1.3.9. Congo Red Assays:

E. coli pLysS strain carrying the PleD* protein and its mutants were grown to a mid-log phase at 37°C and the OD was adjusted to 0.5$_{600nm}$. 2xTY agar plates containing Congo Red at a final concentration of 50 ug/mL and with or without IPTG were spotted with 2.5 ul of the cultures and incubated at 30°C for 24 hours.

1.3.10. High performance liquid chromatography (HPLC):

Enzymatic reactions were performed at 30 ºC with 0.125 µmol/L (PleD* dimer) or 2.5 µmol/L (XAC0610$_{35-880}$ dimer) purified protein in a volume reaction of 100 ul containing 20 mmol/L Tris-Cl pH 8.00, 5 mmol/L MgCl$_2$, 100 mmol/L NaCl, 1 mmol/L or 100 µmol/L of substrate (as indicated). The reactions were quenched at 1 hour or 16 hours adding 900 ul of stop solution (10 mmol/L HCl). The nucleotide products were analyzed by HPLC using a LC-10ADvp Shimadzu equipment with a C-18 column (Jupiter 5u, Phenomenex, 250 mm x 4.6 mm) coupled to a C-18 precolumn (Phenomenex, of 4 mm x 3 mm). The column was equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mmol/L ammonium formate, pH 7). The separation protocol was 2 % of buffer A up to 7 minutes, then the concentration of B was elevated up to 80 % within 17 minutes, this concentration was maintained up to 25 minutes, 80 % to 2 % up to 25 minutes and 2 % up to 50 minutes. When necessary, the products were confirmed using LC-MS/MS (Central Analítica, IQ-USP). The LC protocol was the same as described above and the MS/MS stage was done in an Amazon Speed ETD (Bruker Daltonics) mass spectrometer equipped with an ESI interface. The operating parameters were set as follows: drying gas flow rate, 12 L/min; temperature, 300 ºC; nebulizer, 70 p.s.i; capillary voltage, 4500 V. Samples were analyzed in positive mode, and mass spectra data were recorded across the *m/z* range of 50 – 1000.

1.3.11. Specificity enzymatic assays of GGDEF proteins:

Enzymatic reactions with PleD, PleD*, XAC0610$_{35-880}$, their respective mutants, GSU1658, LIC11128 and LIC11131 were performed at 30 ºC with purified protein (concentrations are indicated by each experiment in results and discussion section) in a reaction volume of 100 ul of reaction buffer (20 mmol/L Tris-Cl pH 8.00, 5 mmol/L $MgCl_2$, 100 mmol/L NaCl) and 1 mmol/L of each tested substrate (ATP, GTP, ITP, 2`dATP, 2`dGTP or 2`dITP). The reactions were diluted and stopped after 1 hour by adding 900 ul of the reaction buffer and stored at -20 ºC until analysis by HPLC-ESI-MS/MS. The products were analyzed by HPLC-ESI-MS/MS in a LC-10ADvp Shimadzu equipment using a C-18 column (Jupiter 5u, Phenomenex, 250 mm x 4.6 mm) coupled to a C-18 precolumn (Phenomenex, 4 mm x 3 mm). The column was equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mmol/L ammonium formiate, pH 7). The separation protocol was 2 % of buffer A up to 7 minutes, then the concentration of B was elevated up to 80 % within 17 minutes, this concentration was maintained up to 25 minutes, 80 % to 2 % up to 25 minutes and 2 % up to 50 minutes. The separated components were detected with an Amazon Speed ETD (Bruker Daltonics) mass spectrometer equipped with an ESI interface. The operating parameters were set as follows: drying gas flow rate, 12 L/min; temperature, 300 ºC; nebulizer, 70 psi; HV, 4500 V. The samples were analyzed in positive mode and mass spectra data recorded across the m/z range of 50-1000. The reference mass for cdiGMP was 691 +-0.2 m/z, for cdiAMP was 659 +-0.2 m/z, for cdiIMP was 661 +-0.2 m/z, for cdidGMP was 659 +-0.2 m/z, for cdidAMP was 627.1 +-0.2 m/z and for cdidIMP was 629.1 +-0.2 m/z.

1.3.12. *In vitro* DGC activity assays of XAC0610 using pyrophosphatase assay:

In vitro activity assays were performed using the EnzChek pyrophosphate kit (Life Technologies) according to the manufacturer's instructions except that the buffer used for the coupled reactions was 100 mmol/L NaCl, 20 mmol/L Tris-HCl pH 8, 5 mmol/L $MgCl_2$. The reactions were initiated by adding GTP or dGTP. Assays were performed in triplicate in Corning Costar 96 well, clear bottomed plates (cat # 3610) containing 0, 100, 200 or 300 µM of 2´dGTP and varying GTP concentrations (0 – 500 µmol/L). Absorbance at 360 nm in each well was measured using a SpectraMax Paradigm plate reader (Molecular devices) using SoftMax Pro 6.2 software.

1.3.13. *In vitro* DGC activity assays of XAC0610 with different ratios of 2´dGTP:GTP:

Activity assays with different ratios of 2´dGTP:GTP were performed using 5 µmol/L XAC0610, GTP 1 mmol/L and different concentrations of 2´dGTP (0 mmol/L, 1 mmol/L, 2 mmol/L, 5 mmol/L and 8 mmol/L) in a reaction buffer 100 mmol/L NaCl, 20 mmol/L

Tris-HCl pH 8.0, 5 mmol/L $MgCl_2$. We performed the assays in duplicate in two different reaction times, 10 and 20 minutes at 30 ºC. The reactions were stopped by warming the samples at 99 ºC for 5 minutes. After that, the samples were diluted 10 times with reaction buffer and analyzed by HPLC-ESI-MS in a LC-10ADvp Shimadzu equipment using a C-18 column (Jupiter 5u, Phenomenex, 250 mm x 4.6 mm) coupled to a C-18 precolumn (Phenomenex, 4 mm x 3 mm). The column was equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mmol/L ammonium formate, pH 7). The separation protocol was 2 % of buffer A up to 7 minutes, then the concentration of B was elevated up to 80 % within 17 minutes, this concentration was maintained up to 25 minutes, 80 % to 2 % up to 25 minutes and 2 % up to 50 minutes. The separated components were detected with an Amazon Speed ETD (Bruker Daltonics) mass spectrometer equipped with an ESI interface. The operating parameters were set as follows: drying gas flow rate, 12 L/min; temperature, 300 ºC; nebulizer, 70 psi; HV, 4500 V. The samples were analyzed in positive mode and mass spectra data recorded across the m/z range of 50-1000. The peaks intensity in the ion chromatogram corresponding to the mass of pppGpdG 853.1 m/z (retention time 8.35 min), cGdGMP 675.1 m/z (retention time 11.2) and cdiGMP 691.1 m/z (retention time 11.9 min) were used to calculate the percentages of each product.

1.3.14. Production of a dGTPase *E. coli* knock-out strain:

A mutant *E. coli* strain from Keio collection (Baba et al., 2006) with a kanamycin resistance cassete inserted in the dGTPase gene was kindly donated by Prof. Beny Spira from the ICB-USP. This *E. coli* strain functioning as a donor strain was grown in LB broth overnight. After that, culture was diluted 1/100 in fresh LB broth supplemented with 10 mmol/L $CaCl_2$ and 0.1 mol/L $MgSO_4$. This culture was grown until OD 600 nm reached 0.3 – 0.5. 200 µl of P1 phages were added to the culture and incubated at 37 ºC for 20 minutes. The supernatant of this culture containing the P1 phages from the donor strain were stored at 4 °C up to its use.

A *E. coli* BL21(DE3) strain was grown in LB broth supplemented with 10 mmol/L $CaCl_2$ and 0.1 M $MgSO_4$. When the culture reached 0.3 – 0.5 OD 600 nm, 100 uL of this culture were mixed with 100 uL of the P1 phages obtained from the donor strain and incubated at 37 ºC for 20 minutes. Then, this misture was plated in fresh medium with 40 mg/mL Kanamycin. Colonies that grown in the plates were subjected to PCR using specific primers (Table 4), to verify that knockout was correctly performed.

Table 4. List of primers used to verify the construction of BL21 Δdgt.

Name	Sequence 5'-3'
Km1	CCGAACTGTTCGCCAGGCTC
Km2	GAGCCTGGCGAACAGTTCGG
Km3	GACTGGGCACAACAGACAATCG
upstream	GCAATGCAGGCCTCAGCGG
downstream	GGAACGGAGAACCTTCCTGGC
dGTP1	CGTCTGGTGCATACATTGATGCG
dGTP2	CGCATCAATGTATGCACCAGACG
dGTP3	GCATCGTCGTTACCGTTCACC

1.3.15. Cyclic dinucleotides extraction from *E. coli* cells:

Cyclic dinucleotides extraction was performed as described (Roy et al., 2013; Teixeira et al., 2018). Briefly, *E. coli* BL21(DE3) cells carrying the corresponding vectors (pBRA_XAC2382, pBRA_XAC2382_HG, pBRA_XAC2382_HGE, pBiocdiGMP-PleD*, pBiocdiAMP-DisA or pBiocGAMP-GSU1658) were grown in 2XTY media overnight. An aliquot of each overnight culture was inoculated in 10 mL of fresh 2XTY media and incubated at 37 ºC. After the bacterial culture reached an OD between 0.6 – 0.8, 1 mmol/L IPTG final concentration was added, and the culture was incubated for 2 hours at 37 ºC at 200 rpm. 1 mL of the culture was transferred to a 1.5 mL Eppendorf tube and centrifuged at 9300 x g for 2 minutes at 4 ºC. The cellular pellet was washed twice with ice-cold PBS (pH 7.2) and resuspended in 100 µl of ice-cold PBS and incubated at 100 ºC for 5 minutes. Ice-cold ethanol was added to a final concentration of 65 % and vortexed for 15 seconds. The sample was centrifuged at 9300 x g for 2 minutes and the supernatant was transferred to a new microcentrifuge tube. This extraction procedure was repeated two more times and all the supernatants were pooled together. The insoluble fraction was retained for subsequent determination of protein content by BCA assays. The pooled ethanol-soluble fraction was vacuum dried after which the pellet was resuspended in 100 ul of distilled water. 40 ul of these samples were subjected to HPLC-ESI-MS analysis using a C-18 column (Jupiter 5u, Phenomenex, 250 mm x 4.6 mm) coupled to a C-18 precolumn (Phenomenex, 4 mm x 3 mm) equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mmol/L ammonium formate, pH 7.0) at 0.6 mL/min. The mixture was eluted using 2 % of buffer B for 7 minutes followed by a 2 % - 80 % B gradient up to 17 min, followed by 80 % B up to 25 min. The separated components were analyzed with an Amazon Speed ETD (Bruker Daltonics) mass spectrometer equipped with an ESI interface (Central Analítica, IQ-USP). The operating parameters were set as follows: drying gas flow rate, 12 L/min; temperature, 300 ºC; nebulizer. 70 psi; HV, 4500 V. The samples were analyzed in positive mode and mass spectra recorded across the m/z range of 50 – 1000. For relative quantification, peak areas of the MS spectrum corresponding to the analyzed cyclic dinucleotide were recorded and normalized with reference to the total protein content of the particular sample.

1.3.16. Analytical gel filtration:

A Superdex 200 Increase 10/300 GL column was calibrated with the protein mixture from the Gel Filtration Calibration Kit MW (SIGMA-ALDRICH cat #MWGF200). The buffer used for column calibration and protein elution was 20 mmol/L Tris-Cl pH 8.0, 100 mmol/L NaCl and 5 mmol/L $MgCl_2$. For column calibration and protein elution the flow rate used was 0.75 mL/min. 100 µl of 50 µM of PleD wild-type or $PleD_{N14Y}$ protein was injected to determine the molecular weight based on the calibration curve.

1.3.17. Multi-angle laser light scattering coupled with size exclusion chromatography:

SEC-MALS analysis was used to determine the molar mass of $XAC0610_{HIS\text{-}35\text{-}880}$ and $XAC0610_{GAF\text{-}PAS1(35\text{-}324)}$. $XAC0610_{HIS\text{-}35\text{-}880}$ molar mass analysis was performed in 20 mmol/L Tris-HCl (pH 8.0), 100 mmol/L NaCl and 5 mmol/L $MgCl_2$ and $XAC0610_{GAF\text{-}PAS1}$ was done in the same buffer but without $MgCl_2$ and in presence or absence of 2 mmol/L cGMP or cAMP. Protein samples (100 µl injection of 30 µM $XAC0610_{HIS\text{-}35\text{-}880}$ or 95 µM $XAC0610_{GAF\text{-}PAS1(35\text{-}324)}$) were separated using a Superdex 200 increase 10/300 GL coupled to a MiniDAWN TREOS multi-angle light scattering system and an Optilab rEX refractive index detector and data analysis was performed using the Astra Software package, version 7.1.1 (Wyatt Technology Corp.).

1.3.18. Negative staining electron microscopy of XAC0610:

Ultra-thin carbon layer on lacey carbon-coated copper grids (TED PELLA, cat #01824) were negatively charged by a glow discharge of 25 seconds at 15 mA. 4.5 ul of purified XAC0610 in 20 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl, 5 mmol/L $MgCl_2$ in different concentrations (50 µg/mL, 100 µg/mL, 200 µg/mL and 300 µg/mL) were placed in the negative charged carbon-coated copper grid for 1 minute. After that, two washes were done with Mili Q water and then stained with 2 % uranyl acetate for 30 seconds. Micrographs were taken in a MET JEOL JEM 2100 transmission electron microscopy with a HAADF detector.

1.3.19. Crystallization tests in 24-well plates:

Crystallization assays were done using the sitting drop-vapor diffusion technique in 24-well plates. Crystallization conditions tested included different commercial kits: Index, Crystal screen, PEG ION screen, PEGRX, Salt RX, Grid Screen Sodium chloride. The procedure uses 0.3 mL of the crystallization buffer in the well reservoir and 1 µl of protein in the pedestal. After this, 1 µl of the crystallization buffer is transferred from the well to the protein solution in the pedestal. The chamber is then sealed with tape and the plates incubated at 18 °C. The crystallization process was observed regularly using a microscope stereoscope.

1.3.20. Crystallization tests in 96-well plates:

Crystallization assays were done using the sitting drop-vapor diffusion technique in 96-well plates. Crystallization conditions tested were those found in the following kits:

Index, Crystal Screen Cryo, PEGRX, PEG/ION screen, SALTRX (Hampton research) and Morpheus (Molecular dimensions). The procedure uses 130 µl of the crystallization buffer in the well reservoir and 2 µl of XAC0610$_{GAF-PAS1(35-324)}$ (in 20 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl) with or without 2 mmol/L cGMP or cAMP in the pedestal. After this, 2 µl of the crystallization buffer is transferred from the well to the protein solution in the pedestal. The chamber is then sealed with tape and the plates incubated at 18 °C. The crystallization process was observed regularly using a microscope stereoscope.

1.3.21. Co-crystallization tests with PleD and different substrates:

Co-crystallization assays were done using the sitting drop-vapor diffusion technique in 24-well plates. Crystallization conditions tested were Glycine 1 M, PEG 20000 13,0 – 16,0 %, Dioxane 1,0 – 2,5 %, pH 8.2 – 9.2. The procedure uses 0.3 mL of the crystallization buffer in the well reservoir and 1 ul of PleDwt (in 20 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl and 5 mmol/L $MgCl_2$) with 0,2 mmol/L cdiGMP and 1 mmol/L substrate (GTP, ATP, ITP, 2´dGTP, 2´dATP, 2´dITP, GTPαS, GpCppp or 3´dGTP) in the pedestal. After this, 1 µl of the crystallization buffer is transferred from the well to the protein solution in the pedestal. The chamber is then sealed with tape and the plates incubated at 18 °C. The crystallization process was observed regularly using a microscope stereoscope.

1.3.22. X-ray diffraction data collection:

X-ray data from the crystals was obtained on the Central Analítica of the Instituto de Química using an X-ray beamline from the rotating anode (MicroMax-007HT) using and R-AXIS IV detector or at the MX2 beamline of the Laboratorio Nacional de Luz Sincrotron, Campinas, Sao Paulo using a Pilatus 2M detector. Crystals were flash-frozen and maintained at 100 K in a stream of a cold nitrogen gas during measurement. Next, diffractions intensities were indexed, integrated, and scaled using iMOSFLM (Battye et al., 2011).

1.3.23. Phase estimation and refinements:

The initial phases for PleD$_{N335T}$ and PleD$_{D344N}$ were solved by molecular replacement using Phaser (McCoy et al., 2007) with the PDB ID: 1W25 model of PleD. Structural refinements were performed using REFMAC5 (Murshudov et al., 2011) in the CCP4i package and COOT (Emsley et al., 2010).

1.4.- RESULTS AND DISCUSSION:

1.4.1. Rational design of PleD* mutants to change specificity from GTP to ATP:

PleD* is a constitutively active version of PleD due to its dimerization-independent of phosphorylation (Aldridge et al., 2003; Paul et al., 2007), it contains four-point mutations: A214T and H234P in the REC2 domain, N120T in the REC1 domain, and Y357N in the GGDEF domain. By sequencing, we also detected that apart of the four-point mutations described in Aldridge et al. (2003), one mutation (N14Y) is present near the phosphorylation site of the PleD Rec1 domain (D53). With the aim of rationally engineering a change on the substrate specificity of PleD* from GTP to ATP, we introduced single mutations in residues D344, N335 and L294 (Fig. 14). The mutants were designed based on the assumption that ATP would bind to the active site in a manner similar to that observed for GTP analogs provided they present proper hydrogen bond donors and acceptors in the active site. The N335 residue was changed to threonine (N335T) to permit a hydrogen bond between the N3 of ATP and the hydroxyl group of threonine. Two mutations were introduced at the position of D344 (D344S and D344N); the first in order to change an H-bond acceptor to an H-bond donor to promote bind to ATP through its N1 and the second mutation could simultaneously form hydrogen bonds with both the N1 and N6 groups of ATP. At position L294, we introduced two single mutations, L294Q and L294E, which could potentially form hydrogen bonds with the N6 amine group of ATP (Fig. 14).

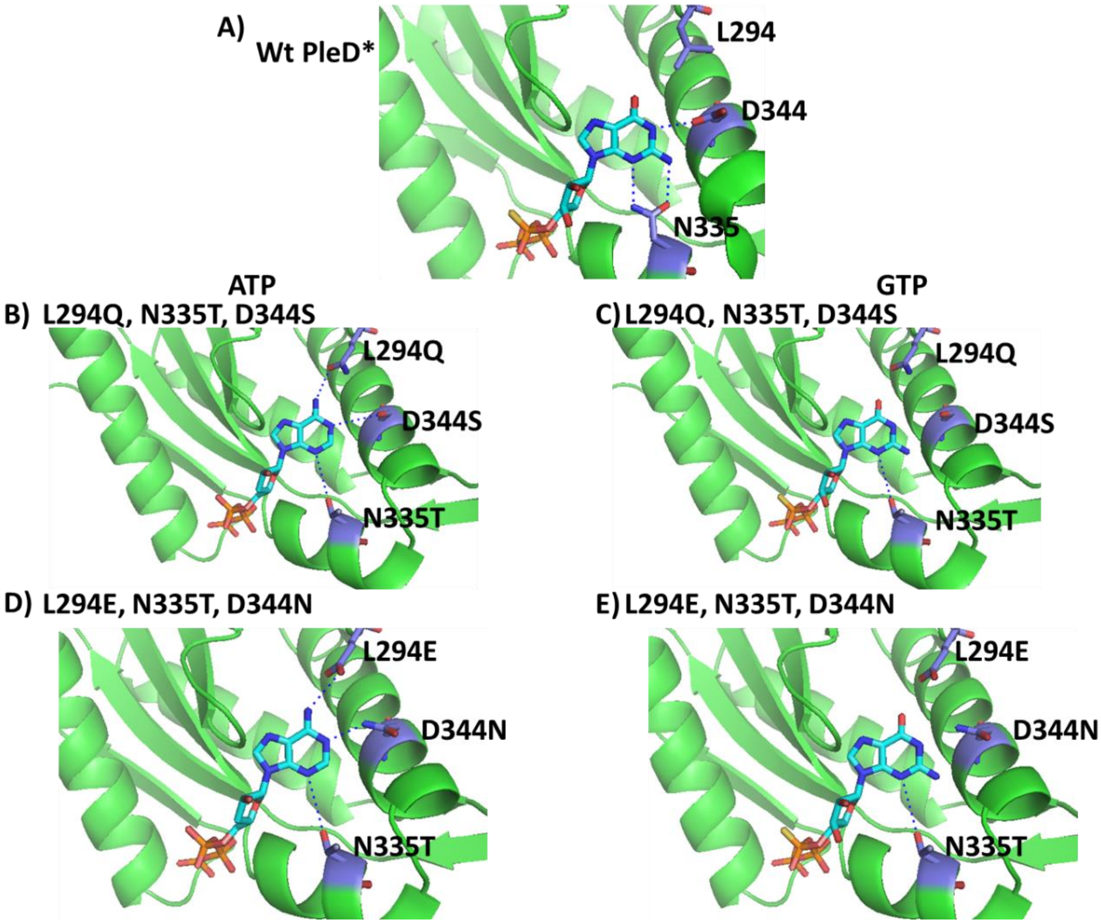


Figure 14: Rationale design of PleD* mutants. A) Wt PleD in complex with GTPαS (PDB ID: 2V0N) showing the residues that interact with guanine moiety. B) and D) Simulation of the interactions of PleD mutants with adenine moiety. C) and E) Simulation of the reduced interactions of PleD mutants with the guanine moiety.

1.4.2. Congo red assays of PleD* mutants:

To rapidly screen for *in vivo* c-di-GMP production by PleD* and its mutants, we performed Congo red assays in *E. coli* expressing the recombinant proteins (Fig. 15). These assays allow us to qualitatively assess the production of intracellular c-di-GMP due the induction of cellulose production. Congo red dye combines with cellulose and changes the colony color from colorless to red depending on the cellulose concentration, which increases in response to elevated c-di-GMP levels. When IPTG was added to induce the expression of PleD* or its mutants, we did not note a visible change of phenotype between the empty vector and PleD* or its mutants. However, we noted that, in contrast to the empty vector, all the colonies with PleD* or its mutants showed grumps. Grumps in liquid and solid media have been associated with c-di-GMP production (Almblad et al., 2021; He et al., 2018; Y. Sun et al., 2020). These results could indicate that these mutations can retain a significant amount of their diguanylate cyclase activity *in vivo* (Fig. 15). It is important to note that the presence of diguanylate cyclase activity in the Congo Red assays does not indicate the absence of other possible enzymatic activities (such as diadenylate cyclase) nor does it provide information regarding the ability of these mutants to bind other substrates (such ATP).

1.4.3. Enzymatic and purification disadvantages using PleD* as a model:

Using the constitutively active version of PleD, PleD*, to perform enzymatic assays presented several problems. First, the protein yield after the affinity and size-exclusion chromatography (see Material and Methods for details) was quite low, forcing us to concentrate several eluted fractions including those that contained protein contaminants. The final preparation of the mutants presented other proteins and the concentration of PleD was not enough to obtain high concentrations in the enzymatic and crystallization assays (Fig. 16A).

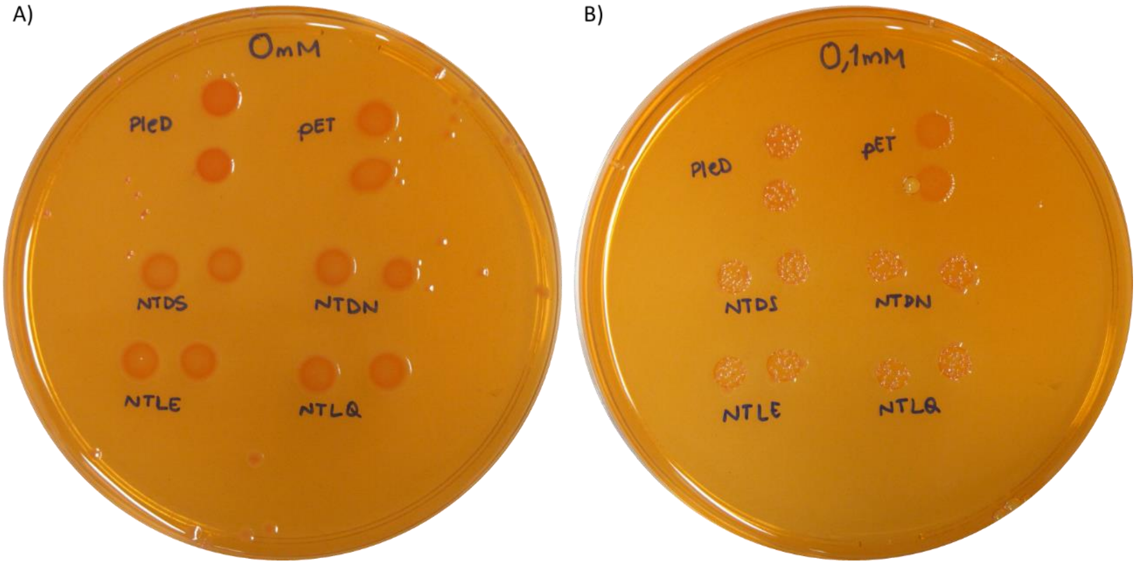


Figure 15. ***In vivo*** **diguanylate cyclase activity of PleD* and its mutants using Congo Red assays.** A: Congo Red assays of *E. coli* cells harboring an expression plasmid of PleD* and its mutants (N335T+D344S (NTDS), N335T+D344N (NTDN), N335T+L294E (NTLE), N335T+L294Q (NTLQ)) in presence of 0 mM IPTG showed colonies very similar to the pET empty vector (pET). B) The same mutants as in A but with 0.1 mM IPTG showing that cells with plasmids harboring PleD* or its mutants have a phenotype of grumped cells.

Secondly, the high affinity of cdiGMP for the inhibition site of PleD (K_i: 0.5 µmol/L; (Wassmann et al., 2007)) leads to the copurification of the feedback-inhibition product with the enzyme, hampering the enzymatic assay. Also, presence of the product in the protein preparation hampers the observation of cdiGMP formation during the enzymatic assays (Fig. 16B).

Finally, the relatively low k_{cat} of PleD (0,054 min^{-1}; (Wassmann et al., 2007)) forced us to employ long incubation times to observe detectable amounts of product by HPLC (Fig.

16C). These results lead us to conclude that PleD* is not a good model for these assays. We therefore looked for a more appropriate diguanylate cyclase to use as a model system.

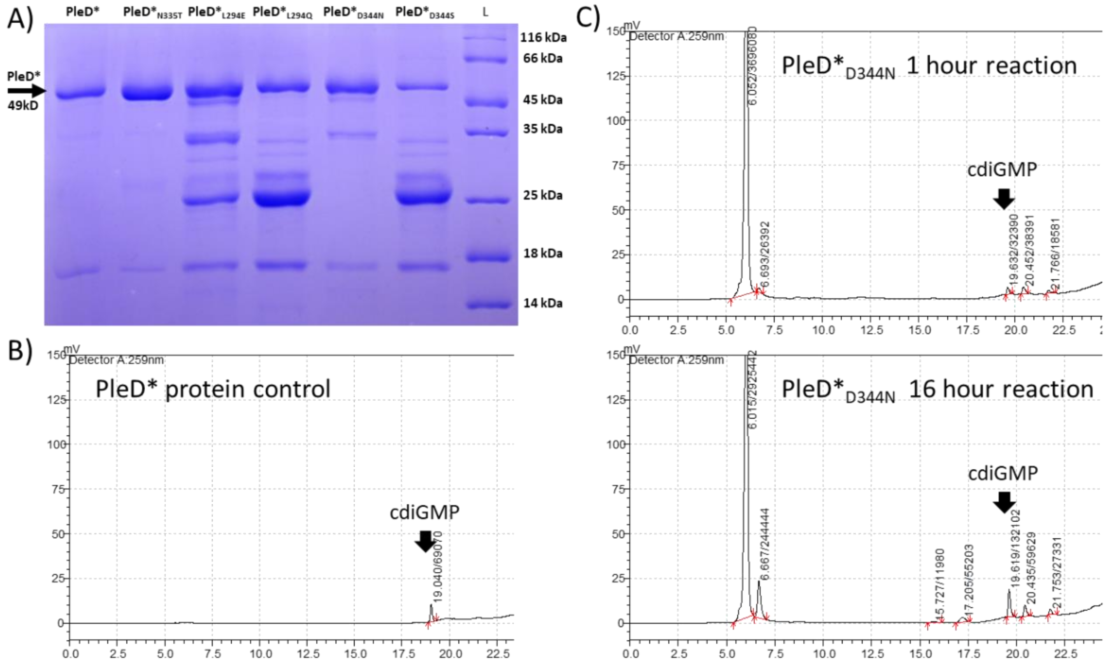
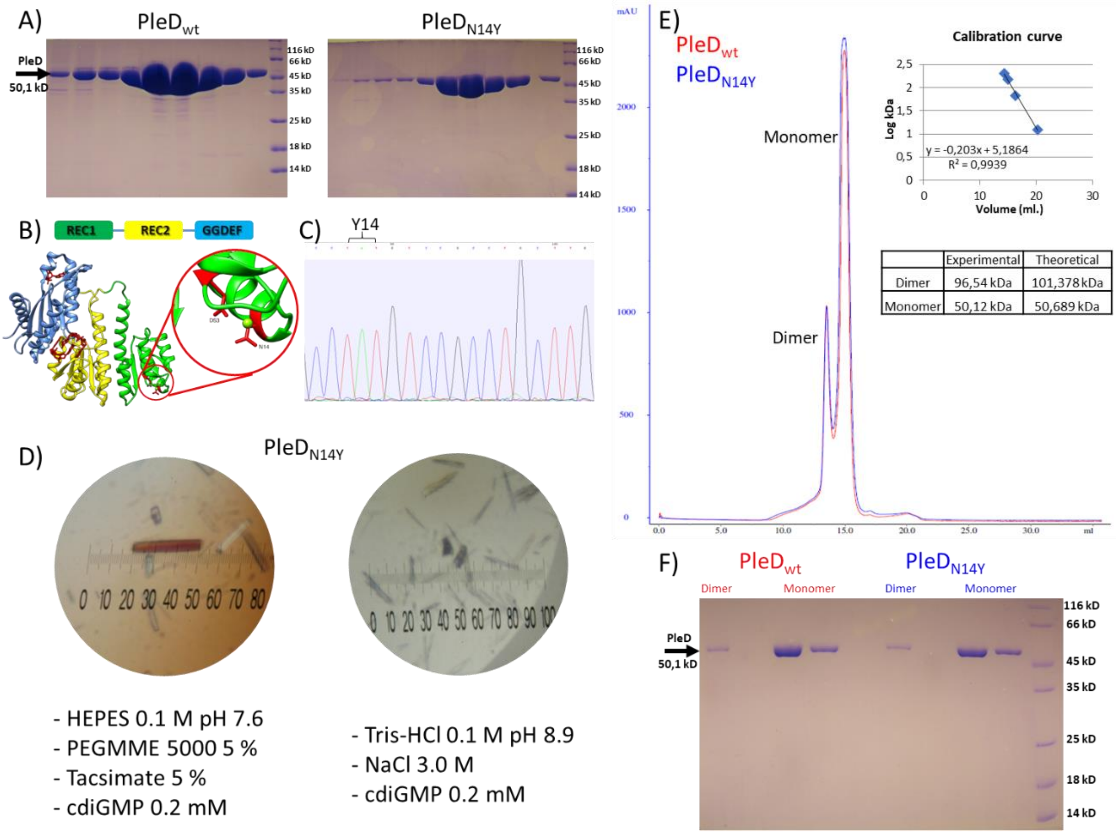


Figure 16. Enzymatic and purification disadvantages using PleD* as a model. A) SDS-PAGE of purified PleD* and its 5 mutants. B) HPLC showing that when we run only a sample of purified protein we identify the presence of cdiGMP. C) Two HPLC results showing the low k_{cat} of PleD*. Above there is a 1-hour reaction and below there is a 16 hours reaction. After 16 hours of reaction the difference in product was very low and other contaminants appear (probably degradation products). Enzymatic reactions were performed at 30 ºC with 0.125 uM (PleD* dimer) protein in 100 ul of 20 mM Tris-Cl pH 8.00, 5 mM $MgCl_2$, 100 mM NaCl, 1 mM of substrate. HPLC was done in a LC-10ADvp Shimadzu using a C-18 column, Jupiter 5u, Phenomenex, of 250 mm x 4,6. The column was equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mM NH_4HCO_2, pH 7). The separation protocol was 2 % of buffer A up to 7 minutes, then B was elevated up to 80 % within 17 minutes and maintained up to 25 minutes.

1.4.4. Comparison between PleD and PleD$_{N14Y}$

As a first alternative to the PleD* mutant, we decided to explore the wild-type PleD protein. We therefore cloned, expressed, and purified PleD from the *Caulobacter* *crescentus* genome (see Methods). As can be seen in Figure 17A, the yield and purity of recombinant PleD is much greater than PleD*. When we analyzed the mutations present in PleD*, we noted that apart of the four-point mutations described in Aldridge et al. (2003), one mutation (N14Y) is present near the phosphorylation site of the PleD Rec1 domain (D53) (Fig. 17B). As described by (Paul et al., 2007), the constitutively active form of PleD* is due its dimerization-independent of phosphorylation.



- HEPES 0.1 M pH 7.6
- PEGMME 5000 5 %
- Tacsimate 5 %
- cdiGMP 0.2 mM

- Tris-HCl 0.1 M pH 8.9
- NaCl 3.0 M
- cdiGMP 0.2 mM

	Experimental	Theoretical
Dimer	96,54 kDa	101,378kDa
Monomer	50,12 kDa	50,689 kDa

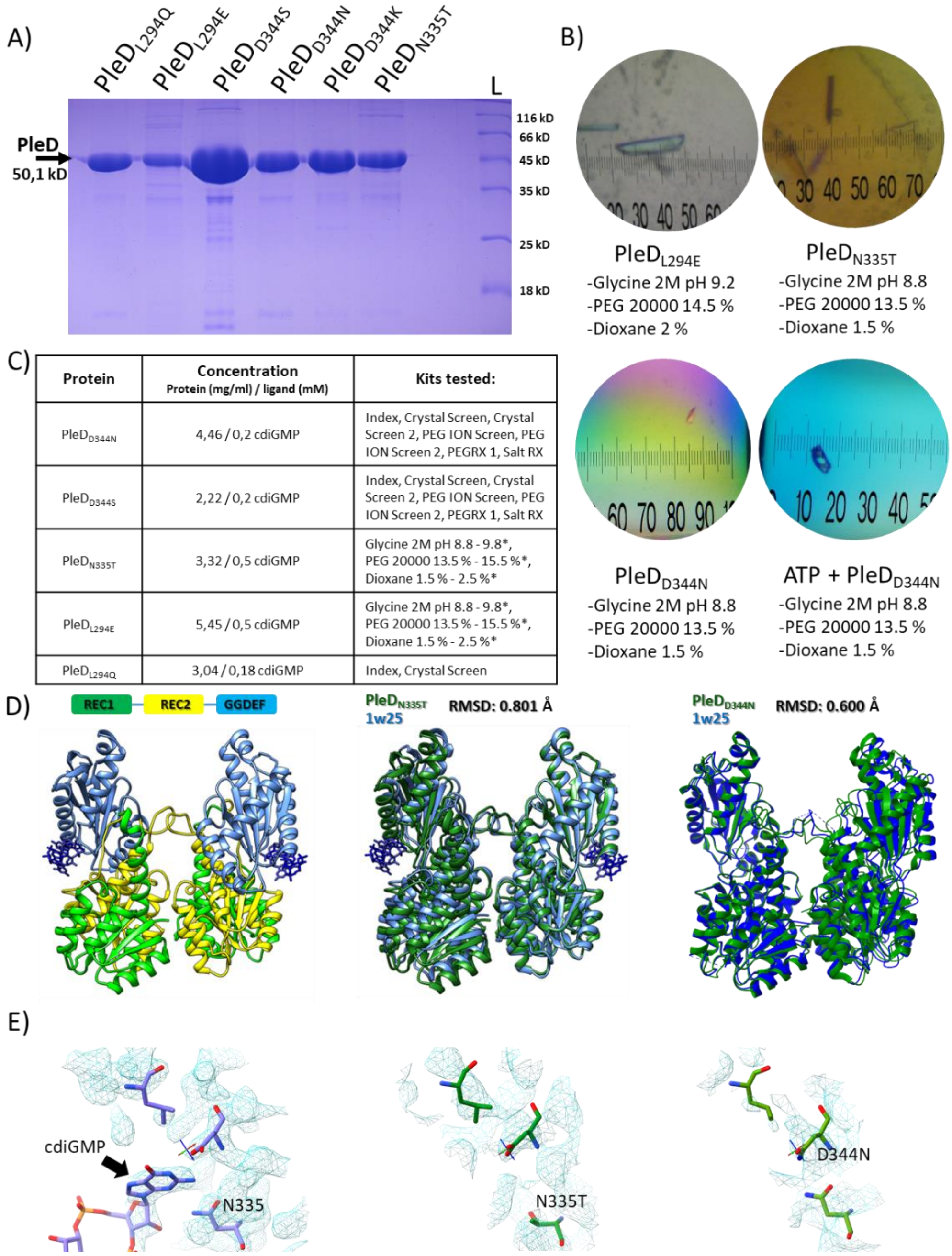
Figure 17. Comparison between PleD and PleDN14Y. A) SDS-PAGE of the purified fractions of PleD (left) and PleDN14Y (right). B) Structural model of PleD (PDB ID: 2V0N) highlighting the position of N14 near to the phosphorylation site (D53). C) Sequencing chromatogram showing the three nucleotides that correspond to the Y14 variation. D) Crystals of PleDN14Y obtained in two different conditions. E) Analytical gel filtration of PleD and PleDN14Y showing the similar oligomerization patterns. F) SDS-PAGE of the peaks showed in E). Superdex 200 Increase 10/300 GL column was calibrated with the protein mixture from the Gel Filtration Calibration Kit MW (SIGMA life Science) (inset). The buffer used was 20 mM Tris-Cl pH 8.0, 100 mM NaCl and 5 mM MgCl2 in a flow rate of 0.75 mL/min. 100 ul of 50 uM protein was injected.

To test if only this mutation (N14Y) could produce a conformational rearrangement producing the constitutively active form of PleD, we introduced this single mutant (Fig. 17C) and analyzed its quaternary structure using analytical gel filtration. As shown in Figure 11E, the oligomerization pattern of PleD and PleD$_{N14Y}$ are very similar, indicating that under the conditions tested, this mutation alone does not result in an increase in the dimerization constant. SDS-PAGE of the fractions eluted from the SEC columns shows that the concentration and purity of the two proteins were very similar (Fig. 17F).

Additionally, because we were able to produce and purify high concentrations of PleD and PleD$_{N14Y}$ we decided to perform crystallization assays. Interestingly, two conditions yielded crystals of the mutant (Fig. 17D) but not of the wild-type protein. This could be due to subtly different conformational states of the two proteins (but not enough to produce a difference in its dimerization pattern). Unfortunately, the PleD$_{N14Y}$ crystals only diffracted up to 10 Å.

1.4.5. Crystallization and diffraction of PleD mutants:

The greater yields and purities obtained with PleD and PleD$_{N14Y}$ opened up the possibility to produce mutants in the guanine recognition pocket of the active site to test their enzymatic activities and to obtain structural models of the mutants by crystallography. However, despite the improved purity and yield of the proteins, the other PleD disadvantage described above (copurification of the feedback-inhibition product with the enzyme and low k_{cat} of PleD) hamper us to use PleD for enzymatic assays. On the other hand, our high protein yields (Fig. 18A) allowed us to continue with the crystallization tests.



Protein	Concentration Protein (mg/ml) / ligand (mM)	Kits tested:
$PleD_{D344N}$	4,46 / 0,2 cdiGMP	Index, Crystal Screen, Crystal Screen 2, PEG ION Screen, PEG ION Screen 2, PEGRX 1, Salt RX
$PleD_{D344S}$	2,22 / 0,2 cdiGMP	Index, Crystal Screen, Crystal Screen 2, PEG ION Screen, PEG ION Screen 2, PEGRX 1, Salt RX
$PleD_{N335T}$	3,32 / 0,5 cdiGMP	Glycine 2M pH 8.8 - 9.8*, PEG 20000 13.5 % - 15.5 %*, Dioxane 1.5 % - 2.5 %*
$PleD_{L294E}$	5,45 / 0,5 cdiGMP	Glycine 2M pH 8.8 - 9.8*, PEG 20000 13.5 % - 15.5 %*, Dioxane 1.5 % - 2.5 %*
$PleD_{L294Q}$	3,04 / 0,18 cdiGMP	Index, Crystal Screen

Figure 18. Crystallization tests of PleD mutants. A) SDS-PAGE of 6 PleD mutants. B) Representative crystals of PleD mutants and the conditions where they crystallized. C) Table showing the concentration of protein and cdiGMP, and kits or conditions tested for each PleD mutant. D) Left: Model of PleDN335T showing a dimer and colored to difference its domains (R-free: 0.289, R-work: 0.217). Middle and Right: Superposition of the structural model of PleDwt (PDB ID: 1w25) and model of PleDN335T or PleDD344N. G) Comparison between the electron density maps on the recognition site of the guanine moiety of PleDwt (right), PleDN335T (middle), and PleDD344N (left).

Figure 18C shows a summary of the kits and crystallization conditions tested for each PleD mutant. The condition with glycine 2M pH 8.8 - 9.8, dioxane 1.5-2.5 % and PEG 20000 13.5-15.5 % and 0.5 mmol/L cdiGMP (similar to that described in (Chan et al., 2004)) produced crystals of three mutants (L294E, N335T and D344N) (Fig. 18B). Furthermore, PleD mutant D344N crystallized in presence and absence of ATP.

Crystals of PleD$_{N335T}$ were analyzed in the MX2 beamline of the LNLS-CNPEM (Campinas) obtaining diffraction patterns with reflections up to 2.5 A. Crystals of the D344N mutant in the absence of ATP diffracted up to 3 A using the rotating anode X-ray source at the Institute of Chemistry (USP). We still have not tried the D344N mutant crystals in the presence of ATP nor the L294E mutant. Using wild type PleD (PDB ID: 1W25) as a model for molecular replacement to calculate initial phases, we were able to calculate an electron density map of these two mutants (Table 5). Figure 12D show PleD$_{N335T}$ and PleD$_{D344N}$ models.

Table 5. Crystallographic data collection and refinement statistics.

Model	PleDN335T	PleDD344N
Data Collection		
Beamline	MX2	MicroMax-007HT
Wavelength (Å)	1.4601	1.5418
Space group	P212121	P212121

Cell axes (Å)	a = 80.98
	b = 85.26
	c = 324.79
		a = 82.34
		b = 85.53
		c = 327.84

Resolution[a]	47.578 – 2.785 (2.95 – 2.78)	48.053 – 2.794 (2.96 – 2.79)
Unique reflections[a]	48191 (7501)	108329 (17776)
Redundancy	3.35 (3.21)	2.56 (2.52)
R_{merge} (%)[a]	10.8 (48)	10.8 (60)
I/Sigma(I)[a]	8.60 (1.50)	10.38 (1.95)

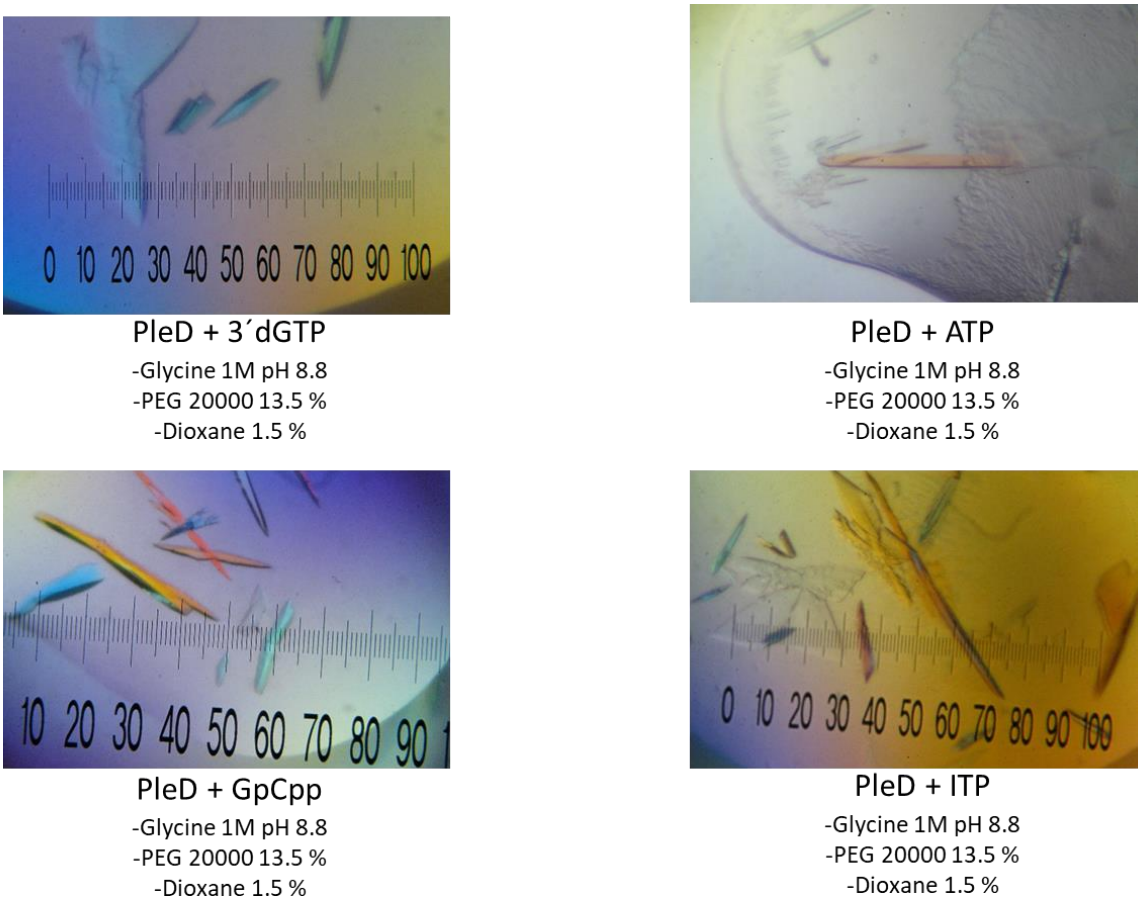
Completeness (%)[a]	84.0 (82.2)	97.7 (98.9)
Refinement		
R_{work}/R_{free} (%)	0.2869/0.3549	0.2401/0.3159
rmsd		
Bond length (Å)	0.012	0.009
Bond angles (°)	1.384	1.193
Ramachandran statistics (%)		
Favored regions	92.81	92.29
Allowed regions	4.92	6.1
Disallowed regions	2.27	1.61

[a] Values in parentheses refer to the highest resolution bin.

The conformational states in these two models are very similar to the non-phosphorylated form of PleD described by (Chan et al., 2004) (Fig. 18D) where two monomers are in contact via their REC domains to form a dimer and two cdiGMP are bound to the inhibition sites of each GGDEF domain crosslinking this domain with the Rec2 domain producing a catalytically non-competent dimer. Interestingly, although the active site is very similar in PleD, $\text{PleD}_{\text{N335T}}$, and $\text{PleD}_{\text{D344N}}$ (Fig. 18E) we did not see any electron density that could correspond to a guanine moiety of cdiGMP indicating that likely these mutations abolished GTP or cdiGMP binding (Fig. 18G). $\text{PleD}_{\text{L294E}}$ diffract just up to 10 A resolution and the other mutants did not form crystals in the tested conditions.

1.4.6. Co-Crystallization of PleD with different substrates:

At the moment there are GGDEF structural models with GTPαS present in the active site (Wassmann et al., 2007; Zähringer et al., 2013). The GTPαS in the active site of these structures does not represent a productive substrate conformation due to significant differences from the conformation that the GMP moiety attains in the cyclic product (Schirmer, 2016).



PleD + 3´dGTP
-Glycine 1M pH 8.8
-PEG 20000 13.5 %
-Dioxane 1.5 %

PleD + ATP
-Glycine 1M pH 8.8
-PEG 20000 13.5 %
-Dioxane 1.5 %

PleD + GpCpp
-Glycine 1M pH 8.8
-PEG 20000 13.5 %
-Dioxane 1.5 %

PleD + ITP
-Glycine 1M pH 8.8
-PEG 20000 13.5 %
-Dioxane 1.5 %

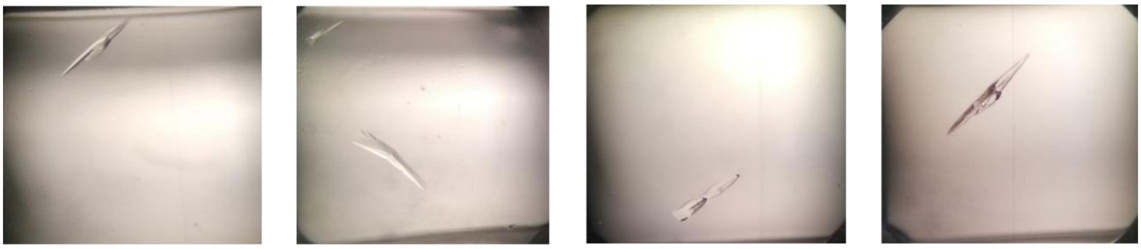
Figure 19. Co-crystallization of PleD with different substrates: 3,5 mg/ml of PleDwt in 20 mM ris-HCl pH 8.0, 100 mM NaCl and 5 mM MgCl2 was mixed with 0,2 mM cdiGMP and 1 mM final concentration of ATP, ITP, dATP, 3´ dGTP or GpCpp. These solutions were used f crystallization assays using the sitting drop-vapor diffusion method. Each photo shows below the substrate used in the crystallization assay and the condition that produce crystals.

In an attempt to obtain structural models that could show a better productive substrate conformation we performed co-crystallization assays of PleD with different GTP analogs including 3´dGTP and GpCpp and other nucleotide triphosphates such as ATP, ITP, and 2´dATP. We obtained PleD crystals in the presence of four of the five molecules tested (Fig. 1); 3´dGTP, GpCpp, ATP and ITP in the condition of glycine 1 M pH 8.8, PEG 20000

13.5 % and dioxane 1.5 %. We brought these crystals to the MX2 line at the LNLS-CNPEM of Campinas but unfortunately these crystals diffracted to no better than 8 A resolution.

1.4.7. Attempts to obtain structures of GGDEF domain in complex with linear products:

To elucidate how a GGDEF domain could bind a linear product in its active site that could represent an intermediate state of the reaction, we attempted to cocrystallize PleD in the presence of GTP and nonhydrolyzable GTP analogs (GTPαS, 3´dGTP, GpCpp). We obtained small crystals of PleD in presence of GTP and 3´dGTP, GTP and 2´dGTP, GTP and GpCpp (Fig. 20). However, X-ray diffraction experiments were not carried out.



PleD + GTP + 2´dGTP
-Glycine 1M pH 9.5
-PEG 20000 13.5 %
-Dioxane 1.5 %

PleD + GTP + 3´dGTP
-Glycine 1M pH 9
-PEG 20000 15 %
-Dioxane 2.5 %

PleD + GTP + 3´dGTP
-Glycine 1M pH 9
-PEG 20000 13.5 %
-Dioxane 2.5 %
-cdiGMP 0.2 mM

PleD + GTP + GpCpp
-Glycine 1M pH 9.5
-PEG 20000 15 %
-Dioxane 1.5 %
-cdiGMP 0.2 mM

Figure 20: PleD crystallizes in presence of GTP and non-hydrolyzable GTP analogs. Solutions of 16 mg/ml of PleDwt in 20 mM Tris-HCl pH 8.0, 100 mM NaCl, 5 mM MgCl2, 1mM GTP with or without 0,2 mM cdiGMP and 1 mM final concentration of 2´dGTP, 3´ dGTP or GpCpp were prepared. These solutions were used for crystallization assays using the sitting drop-vapor diffusion method. Each photo shows below the combination of ligands used in the crystallization assay and the condition that produce crystals.

1.4.8. RNA-based biosensors to detect cyclic dinucleotides *in vivo*:

RNA-based biosensors are RNA molecules able to recognize a target molecule and emit fluorescence upon binding (Fig. 21A). Ming Chen Hammond´s group (Kellenberger et al., 2013; Kellenberger, Chen, et al., 2015; Kellenberger, Wilson, et al., 2015) in University of Berkeley developed a set of these biosensors able to specifically recognize three cyclic dinucleotides (cdiGMP, cdiAMP and cGAMP). These biosensors are embedded in a tRNA scaffold that stabilizes the whole RNA biosensor (Fig. 21D).

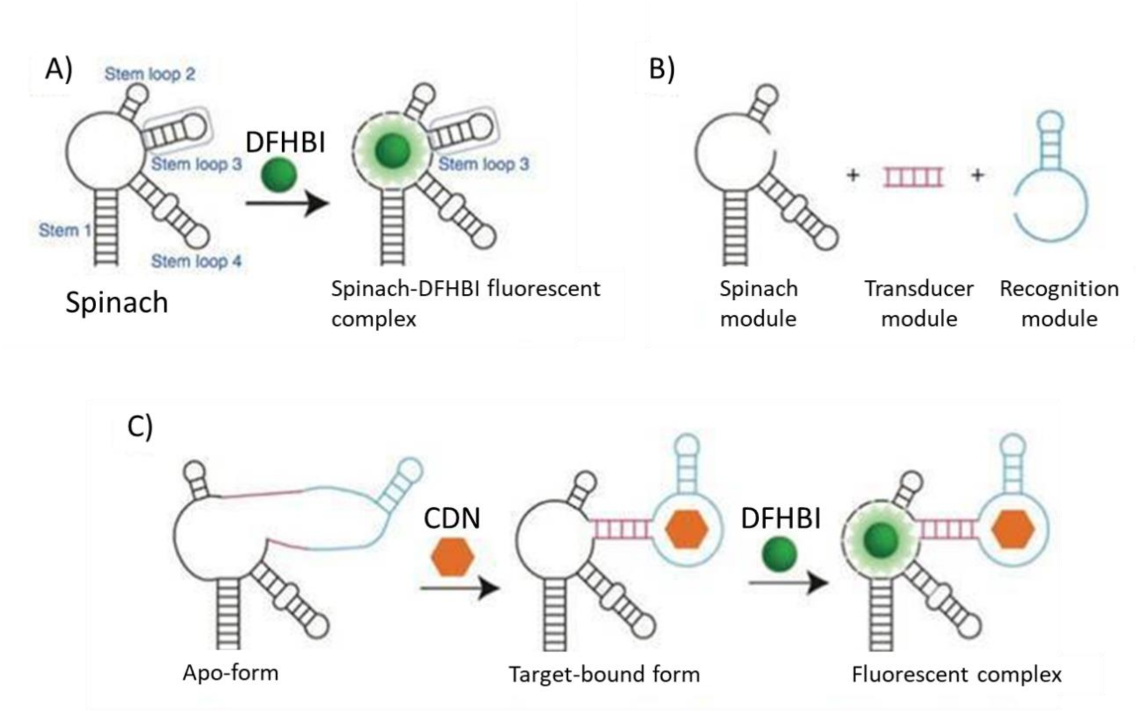


Figure 21: RNA-based biosensors (modified from Strack et al. 2014 and Kellenberger et al. 2015): A) Spinach aptamer is capable to recognize DFHBI; DFHBI increases its fluorescence when binds to Spinach aptamer. B) Construction of a biosensor replace one stem loop in the spinach aptamer for a transducer module connected to a recognition module (aptamer able to recognize the target molecule). C) In absence of the target molecule (Apo-form) neither the recognition module nor the transducer module stabilizes; when target molecule binds to the recognition module transducer module is stabilized allowing the binding of the DFHBI to the Spinach aptamer that results in increasing fluorescence. D) Example of a biosensor to detect cdiGMP composed for a tRNA scaffold, the spinach aptamer and the cdiGMP aptamer.

The biosensor contains a spinach aptamer able to bind the fluorescent DFHBI molecule (spinach module) connected by a double stranded RNA (transducer module) to the cyclic dinucleotide aptamer that binds the target molecule (recognition module) (Fig. 21B and 21D). When the biosensor is intracellularly expressed, binding of the target molecule

(CDN) stabilizes the spinach module that consequently binds the DFHBI molecule increasing its fluorescence (Fig. 21C). Thus, fluorescence intensity is correlated to the intracellular concentration of the corresponding cyclic dinucleotide.

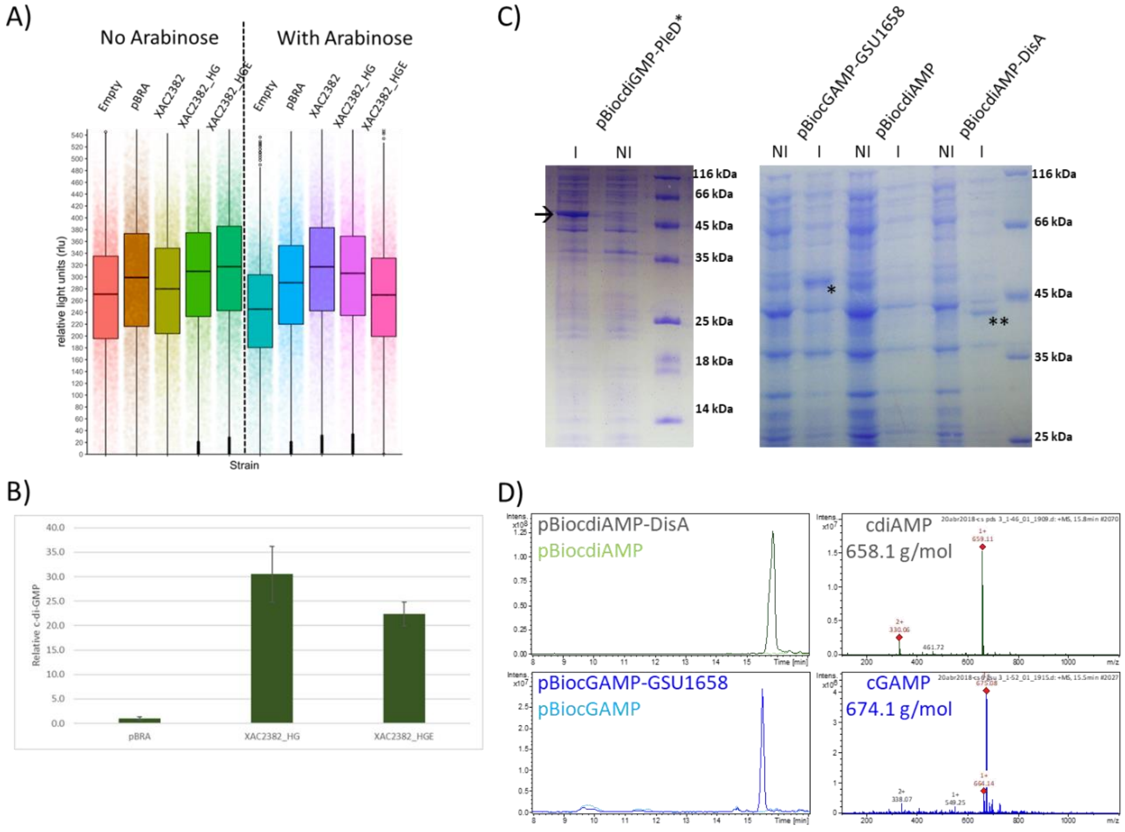


Figure 22: Tests with RNA-based biosensors to detect Cyclic Dinucleotides: A) Fluorescence microscope assays using *E. coli* (BL21DE3) strains overexpressing the diguanylate cyclases XAC2382 did not show the expected increase in fluorescence when compared with strains without overexpression of diguanylate cyclases. B) To confirm that strains overexpressing XAC2382 had more cdiGMP concentration than strains with empty vector (pBRA) we did cdiGMP extraction procedures from cells. We observed a ~30 fold or ~20 fold increasing in cdiGMP concentrations in cells overexpressing XAC2382_HG or XAC2382_HGE respectively. C) Expression assays of biosensor plasmids (pBiocdiGMP-PleD*, pBiocGAMP-GSU1658 and pBiocdiAMP-DisA) showed overexpression of the cyclases (→ = PleD* (50.5 kDa), * = GSU1658 (54.4 kDa), ** = DisA (44.0 kDa)) in *E. coli* BL21(DE3). D) Cyclic dinucleotide extractions from *E. coli* BL21(DE3) transformed with pBiocdiAMP-DisA or pBiocGAMP-GSU1658 showed higher intracelular concentrations of cdiAMP or cGAMP than strains transformed with pBiocdiAMP or pBiocGAMP.

The biosensors sequences to detect three different cyclic dinucleotides (cdiGMP, cdiAMP and cGAMP) under the control of the T7 promoter (see Methods) were cloned in the SphI and BglII restriction sites of pET-28a obtaining the plasmids named pBiocdiGMP, pBiocdiAMP, and pBiocGAMP. To test the cdiGMP biosensor, we transformed *E. coli* BL21(DE3) with the biosensor plasmid (pBiocdiGMP) and with two different constructions of the diguanylate cyclase XAC2382 in the pBRA plasmid under the control of the arabinose promoter (pBRA_XAC2382_HG and pBRA_XAC2382_HGE).

These strains were used to perform fluorescence microscopy assays and cyclic dinucleotide extractions from cells. After induction by addition of IPTG and arabinose we were not able to detect differences between fluorescence of strains with the empty vector (pBRA) or with XAC2382 constructions (pBRA_XAC2382_HG or pBRA_XAC2382_HGE) (Fig. 16A). To verify if strains overexpressing diguanylate cyclases in fact presented elevated intracellular concentrations of cdiGMP, we performed cyclic dinucleotide extractions from cells followed by HPLC-MS quantification. These results showed that strains overexpressing the two XAC2382 constructs XAC2382_HG and XAC2382_HGE, have ~30 fold or ~20 fold more cdiGMP, respectively, than the strain with the empty vector (pBRA) (Fig. 22B). We conclude that the biosensor is either: i) not expressing, ii) is not correctly functional or iii) is saturated with cdiGMP under all conditions tested.

To test the functionality of the other plasmid biosensors (pBiocGAMP and pBiocdiAMP), we cloned the protein GSU1658 (known to have guanylate-adenylate cyclase activity) under the control of the T7 promoter of the pET28a plasmid with the cGAMP biosensor (pBiocGAMP) producing the plasmid pBiocGAMP-GSU1568. Similarly, DisA (known to have diadenylate cyclase activity) was cloned into the pET28a plasmid with the cdiAMP biosensor (pBiocdiAMP) to produce the pBiocdiAMP-DisA plasmid. We analyzed the expression of the proteins by SDS-PAGE and verified that after IPTG addition protein bands with the expected sizes appear (PleD* = 50.5 kDa, GSU1658 = 54.4 kDa, DisA =44.0 kDa) (Fig. 22C).

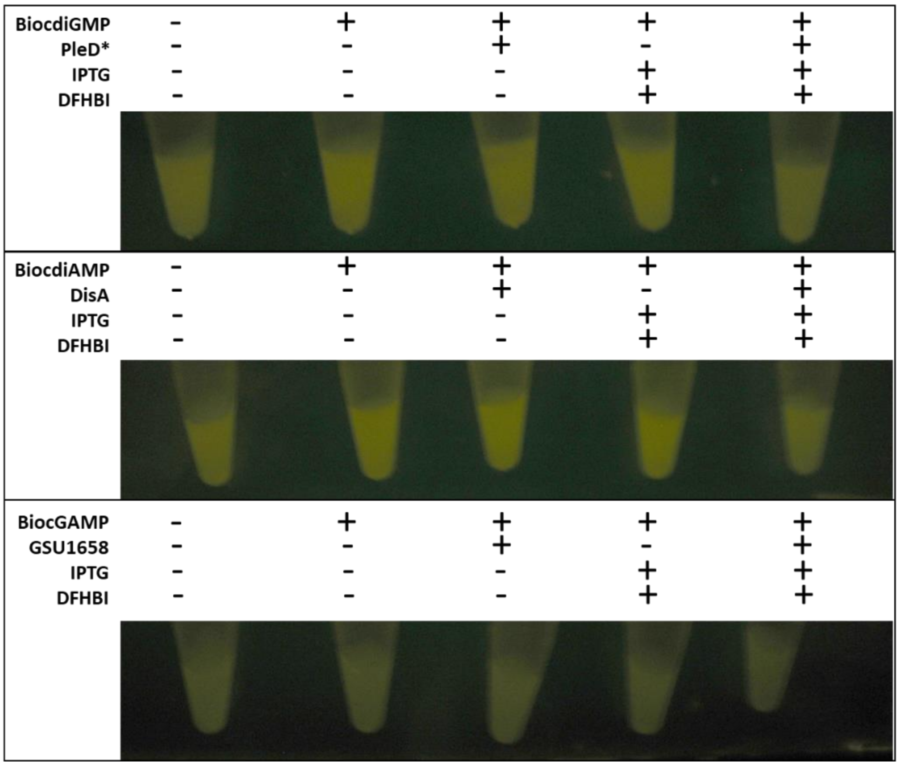


Figure 23: Tests with RNA-based biosensors to detect Cyclic Dinucleotides: Microcentrifuge tubes with BL21(DE3) *E. coli* cells carrying the biosensor plasmids (pBiocdiGMP, pBiocdiAMP, or pBiocGAMP) with or without the corresponding cyclic dinucleotide synthethases (PleD*, DisA, or GSU1658). Induction of the biosensor and the synthethases were done by adding IPTG 0.1 M. DFHBI was added at a final concentration of 50 uM. The tubes where observed in a blue-light transiluminator with an emission range of 445 – 505 nm (excitation range of DFHBI is 400 – 500 nm).

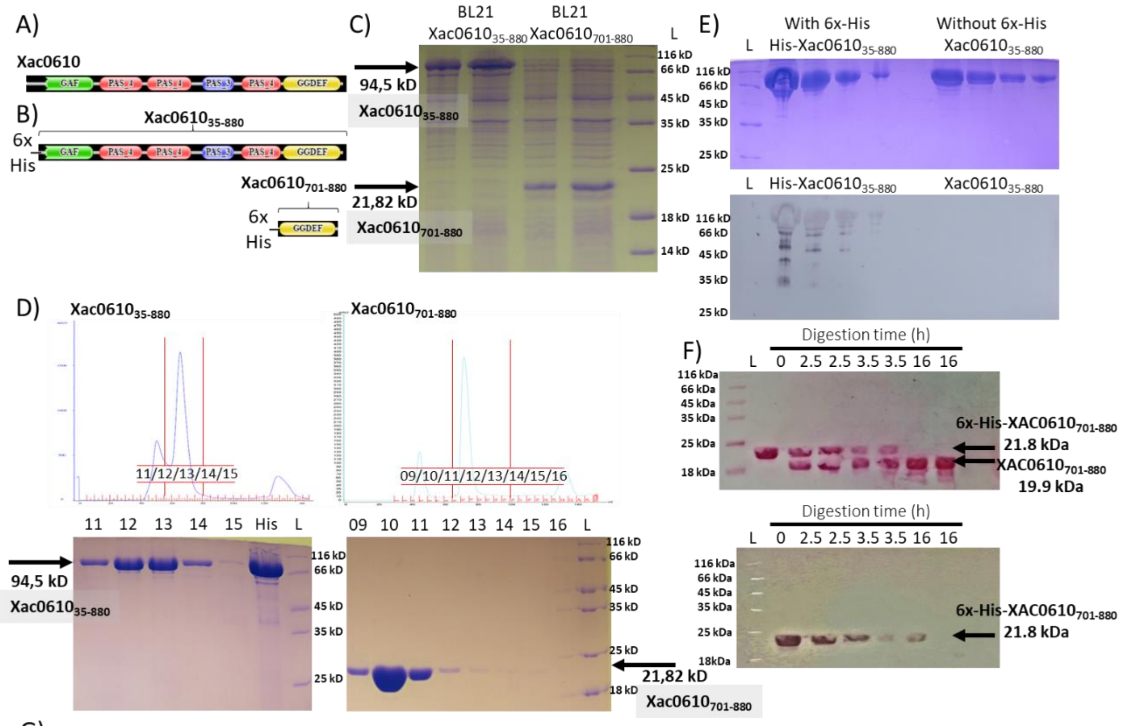
To determine if intracellular cyclic dinucleotide concentration increased in these cells, we performed cyclic dinucleotide extractions in *E. coli* BL21(DE3) cells containing pBiocdiAMP, pBiocdiAMP-DisA, pBiocGAMP and pBiocGAMP-GSU1658. Results of these experiments showed that cells overexpressing GSU1658 or DisA present higher concentrations of the corresponding cyclic dinucleotide (cGAMP or cdiAMP) than cells with pBiocGAMP or pBiocdiAMP plasmids indicating that these enzymes are active in the cells (Fig. 22D).

However, we did not observe differences in the fluorescence neither between the tubes with cells expressing or not the dinucleotide synthetases (PleD*, DisA, or GSU1658) nor with or without DFHBI (Fig. 23).

Taken together, these results show that biosensor plasmids are not functioning as expected. In the case of pBiocGAMP and pBiocdiAMP, microscopy analysis should be carried out to test if this procedure with more sensitivity could detect the production of cyclic dinucleotides. In all the cases, the inability to detect cyclic nucleotides production based on the difference of the fluorescence could be due to incorrect cloning of the biosensors in the plasmid (unlikely because sequences in the plasmids were confirmed by sequencing), low mRNA expression of the biosensor or because of low binding capacity of the biosensor (improper folding for example).

1.4.9. Structural studies of XAC0610:

Due to the low enzymatic activity and product inhibition displayed by PleD, we decided to work with the XAC0610 DGC from *Xanthomonas citri*. XAC0610 is a protein with five annotated domains (one GAF, four PAS and one GGDEF domain) (Fig. 17A). This enzyme has the advantages being highly soluble, a high k_{cat} (65 min^{-1}) and does not exhibit allosteric product inhibition (Oliveira et al., 2015).



G)

With 6x-His

Protein	Kits tested:
His-Xac0610$_{35-880}$	Index, Crystal Screen, Crystal Screen lite, PEGRX 1, PEGRX 2, Grid Screen Sodium Chloride
His-Xac0610$_{701-880}$	Index, Crystal Screen, Crystal Screen 2, PEGRX 1, PEGRX 2, Salt RX 1, Salt RX 2

Without 6x-His

Protein	Kits tested:
Xac0610$_{35-880}$	Index, Index with chimiotrypsin, Index with trypsin, Crystal Screen 1 Crystal Screen 1 with chimiotrypsin, Crystal Screen 1 with trypsin, Crystal Screen 2 Crystal screen 2 with chimiotrypsin, Crystal screen 2 with trypsin
Xac0610$_{701-880}$	Index, Crystal Screen, Crystal Screen 2, PEGRX 1, PEGRX 2, Salt RX 1, Salt RX 2

Figure 17. Induction, purification and crystallization tests of Xac0610. A) Scheme of the XAC0610 architecture showing its 6 domains. B) Scheme of the two constructions used (Xac0610$_{35-880}$ above and Xac0610701-880 below). C) SDS-PAGE showing the induction of the two constructions. D) Size-exclusion chromatography showing the last step of purification of the constructions. E) SDS-PAGE (above) and western blot (below) showing Xac0610$_{35-880}$ before and after removing the His6x-tag. F) SDS-PAGE (above) and western blot (below) showing Xac0610$_{701-880}$ before and after removing the His6x-tag in different time intervals. G) Tables summarizing the crystallization kits tested for each protein.

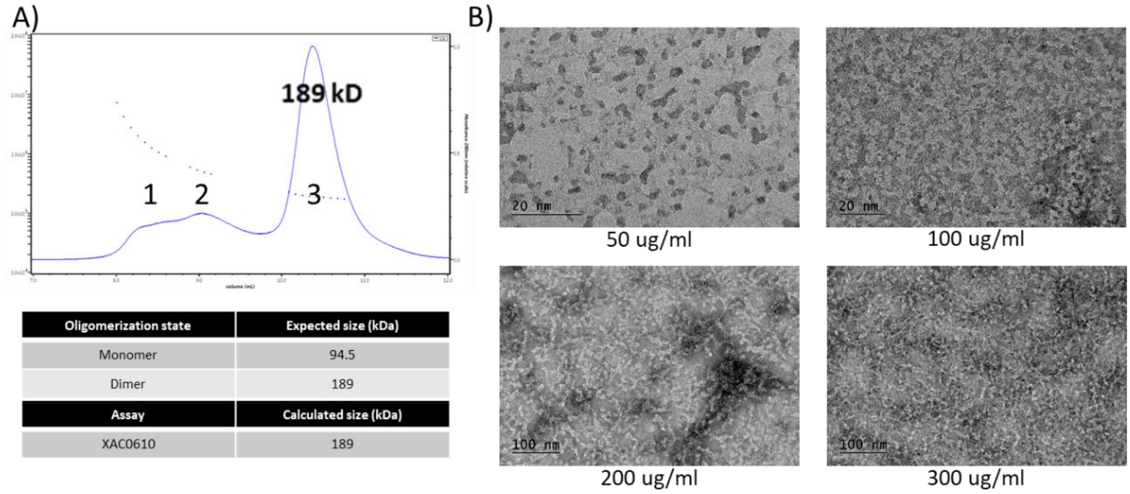
The main disadvantage that we have in using XAC0610 is the absence of a structural model of its GGDEF domain that allows us to rationally design mutants with different specificities. For this reason, we began our work with XAC0610 doing crystallization tests of two constructions (Fig. 24B): One that contains all the five domains (XAC0610$_{35-880}$)

and other that contains only the GGDEF domain ($XAC0610_{701-880}$) both with N-terminal His_{6x}-tag.

Figure 24C is shows induction assays with the two constructions showing the presence of bands with the expected molecular weights ($XAC0610_{35-880}$ = 94.5 kDa, $XAC0610_{701-880}$ = 21.8 kDa). Figure 24D shows the purified proteins obtained after the last step of purification protocol by size exclusion chromatography in a Superdex S200 ($XAC0610_{35-880}$) or Superdex S75 ($XAC0610_{701-880}$) column.

We also performed experiments to remove the His-tag from $XAC0610_{35-880}$ and $XAC0610_{701-880}$. Figure 24E and 24F show that we were able to remove the His-tag in these two constructions and Figure 24G shows the crystallization kits tested for each preparation. Unfortunately, none of the conditions with the three preparations produced protein crystals.

We also carried out SEC-MALS experiments to confirm the oligomerization state of XAC0610, and we found that it elutes primarily as a dimer (Fig. 25A). With a dimer molecular mass of approximately 189 kDa, XAC0610 is near the edge of the mass limit to produce high resolution models using electron microscopy techniques. Thus, we also used negative-staining electron microscopy with four different concentrations of XAC0610.



Oligomerization state	Expected size (kDa)
Monomer	94.5
Dimer	189
Assay	**Calculated size (kDa)**
XAC0610	189

Figure 25: XAC0610 is a small protein for negative staining preparation assays: A) SEC-MALS analysis of XAC0610$_{35\text{-}880}$. A 100 ul protein sample (30 uM) was separated through a Superdex 200 column coupled to a multi-angle light scattering system and refractive index detector. Protein elution was monitored at 280 nm (continuous line). Circles indicate the calculated molecular mass distributions. The major peak (3) had a polydispersity of 1.00 (± 0.5 %) and a measured molar mass of 189 kDa, the same that the expected for a dimer of XAC0610$_{35\text{-}880}$. Peak 1 and 2 showed a molar mass of 1730 kDa and 499 kDa respectively and a polydispersity of 1.27 (± 0.2) and 1.01 (± 0.3). B) Micrographs of electron microscopy of four different concentrations of XAC0610 showing not well-defined particles.

Unfortunately, we did not observe well defined particles in these assays (Fig. 25B), presumably due to the small size and possible flexibility of this protein or the procedure used to prepare the negatively stained grid. We believe that going directly to cryo-EM assays may be more productive than trying to improve the negative staining images.

1.4.10. High-throughput purification assays of XAC0610 constructions:

Because the first two constructions did not produce satisfactory results, we designed 47 different constructions containing one, two, three, four or five domains in all possible combinations of XAC0610. Table 2 presents a list of the 47 constructions with the domain coverage and the first and the last amino acid.

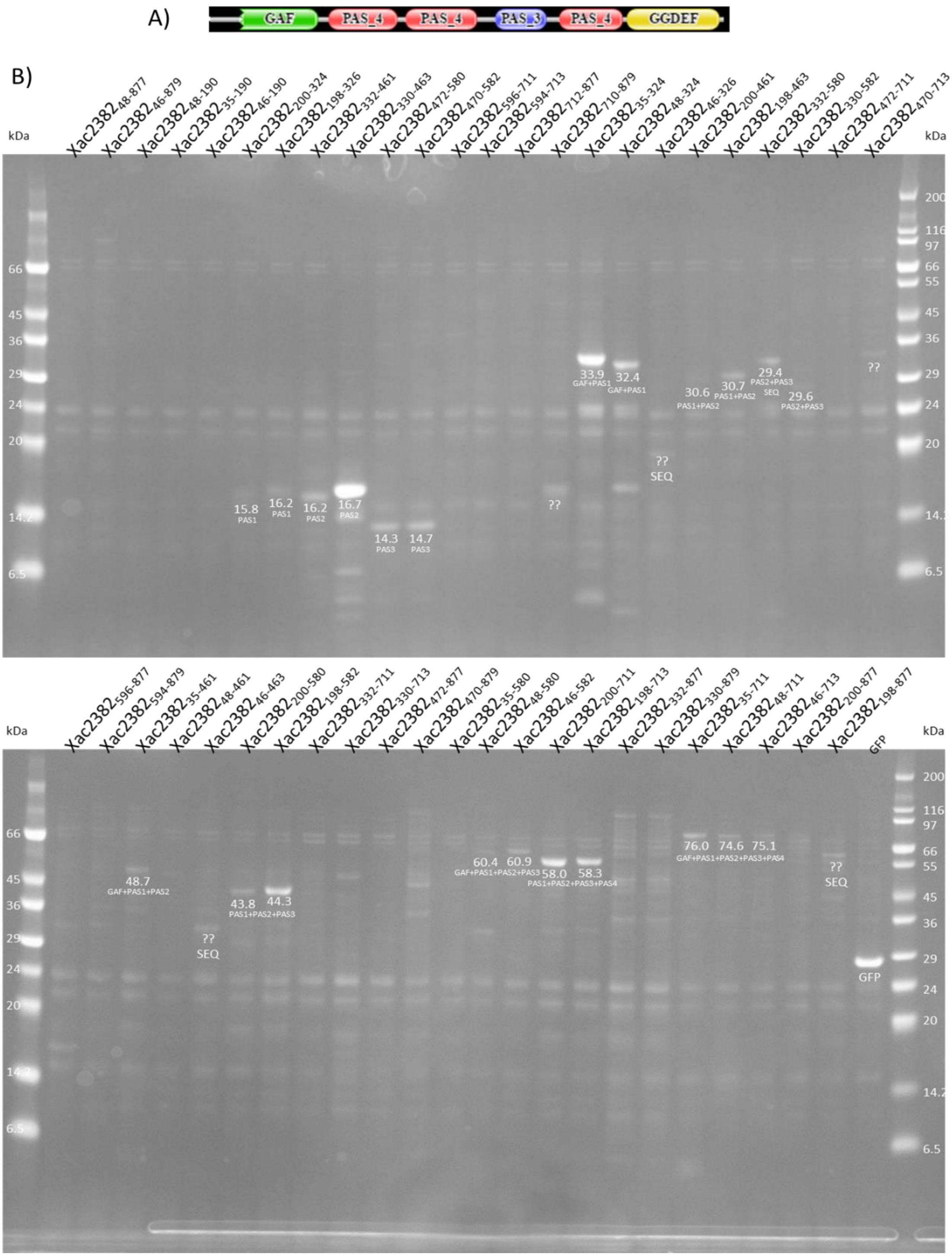


Figure 26. High-throughput purification assays of XAC0610 constructions: A) Domains architecture of XAC0610, B) 47 seven different constructions were cloned in pOPINF vectors and expression assays were done using BL21(DE3) LEMO strain (see materials and methods). Purification assays were done using Ni-NTA magnetic beads and samples were prepared to SDS-PAGE. Expected bands were labeled with its expected molecular weight and the corresponding domain coverage. Other bands that appear in the gel were labeled as "??SEQ" when the expected size was not correct.

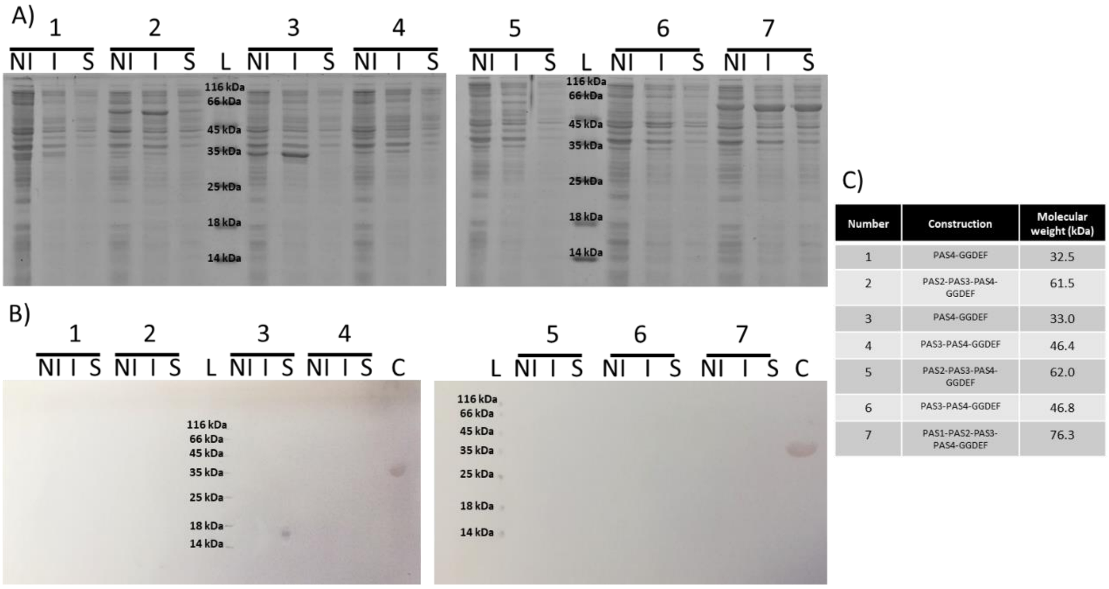
The processes of cloning, expression and purification were done in the Oxford Protein Production Facility (OPPF, Harwell Oxford Campus, UK) following the protocols described in the Methodologies section. Of the 47 seven constructions, 22 were satisfactorily purified by Nickel affinity chromatography (Fig. 26).

PAS1, PAS2 and PAS3 single domains were purified using two different constructions each. Neither GAF, PAS4 nor GGDEF single domains were purified. We were able to purify the GAF domain when linked to the PAS1 domains in two different constructions. PAS1 linked to PAS2 and PAS2 linked to PAS3 were also purified in two different constructions each.

Three constructions containing three consecutive domains were purified (GAF-PAS1-PAS2 and two constructions of PAS1-PAS2-PAS3). In the case of constructions comprising four domains, we were able to purify four different constructions: two for GAF-PAS1-PAS2-PAS3 and two of PAS1-PAS2-PAS3-PAS4. Finally, three different constructions comprising GAF-PAS1-PAS2-PAS3-PAS4 domains were satisfactorily purified and can be used for posterior studies on the structure of XAC0610.

1.4.11. Further purification and analysis of XAC0610 constructs:

We were not able not able to obtain purified samples of constructs containing the XAC0610 GGDEF domain in the high-throughput assays. We therefore attempted to express 7 of the constructions containing GGDEF domains using traditional methods. However, none of these constructions led to detectable expression in *E. coli* BL21 as shown in SDS-PAGE and western blot assays using anti-His (Fig. 27).

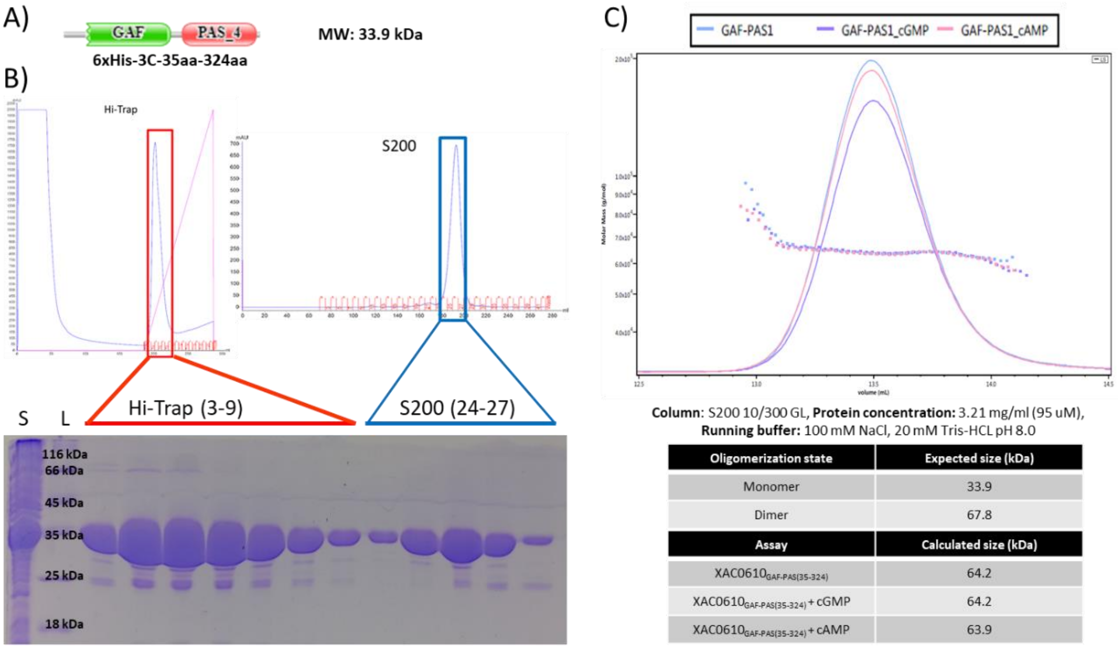


Number	Construction	Molecular weight (kDa)
1	PAS4-GGDEF	32.5
2	PAS2-PAS3-PAS4-GGDEF	61.5
3	PAS4-GGDEF	33.0
4	PAS3-PAS4-GGDEF	46.4
5	PAS2-PAS3-PAS4-GGDEF	62.0
6	PAS3-PAS4-GGDEF	46.8
7	PAS1-PAS2-PAS3-PAS4-GGDEF	76.3

Figure 27: OPPF constructions containing XAC0610 GGDEF domain do not express: pOPINF plasmids containing the domains showed in the panel in the left (C) were used to perform induction assays in *E. coli* BL21 strain. SDS-PAGE (A) and western blot using anti-6xHis (B) experiments of the not induced (NI), induced (I) and soluble fraction of these constructions shows not expression in any of the experiments.

New structural information about the signaling domains of XAC0610 also could help us better understand the physiological role of this protein. Therefore, we began working to try to elucidate the structure of the signaling domains of XAC0610 beginning with a construction containing the first two domains of this protein, GAF and PAS1. We were able to purify this protein (Fig. 28.A) and determined the oligomerization state using SEC-MALS, which revealed that this fragment elutes as a dimer (Fig. 28.B).

After that, we performed crystallization trials with several kits, also testing if the presence of cGMP or cAMP allows this construction to crystallize (since several GAF domains bind cNMPs (Heikaus et al., 2009; Ho et al., 2000; Martinez et al., 2002, 2005; Muradov et al., 2003; Wang et al., 2010; Zoraghi et al., 2004). Unfortunately, we did not obtain crystals.



Column: S200 10/300 GL, **Protein concentration:** 3.21 mg/ml (95 uM), **Running buffer:** 100 mM NaCl, 20 mM Tris-HCL pH 8.0

Oligomerization state	Expected size (kDa)
Monomer	33.9
Dimer	67.8
Assay	**Calculated size (kDa)**
XAC0610$_{GAF-PAS(35-324)}$	64.2
XAC0610$_{GAF-PAS(35-324)}$ + cGMP	64.2
XAC0610$_{GAF-PAS(35-324)}$ + cAMP	63.9

Figure 28: XAC0610$_{GAF-PAS(35-324)}$ is a soluble construction and preferentially form dimers: A) Scheme of XAC0610$_{GAF-PAS(35-324)}$ construction, this construction contains a 6xHis N-terminal tag followed by a HRV 3C protease cleavage sequence before the coding sequence of XAC0610 from aminoacid 35 to 324. The expected mass of this construction is 33.9 kDa. B) The panel on the left show the Nickel affinity chromatogram. Fractions that contain the expected protein were concentrated and a size exclusion chromatography in a S200 column was performed (panel on the right). Fractions containing peaks from the two chromatography were run in an SDS-PAGE (panel below). C) SEC-MALS assays of purified XAC0610$_{GAF-PAS(35-324)}$, with or without 2 mM cGMP or cAMP. The three assays show a peak with a molecular mass corresponding to a dimer.

1.4.12. Enzymatic assays of XAC0610 and its mutants:

We designed and produced five different XAC0610$_{35-880}$ mutants with single amino acid substitution in its GGDEF domain (D771R, D771S, R793H, L721E, L721Q) (D771, L721 and R793 corresponds to PleD residues D344, L294 and R366 respectively) (Fig. 29A). We selected XAC0610$_{35-880}$ construction due to that it contains all the XAC0610 domains and its already showed activity *in vitro* (Oliveira et al., 2015). These mutants were designed based on the homologous PleD structural model in order to attempt to change the enzyme´s specificity from GTP to ATP. All the mutated residues are highly conserved in the GGDEF domains annotated in the database (L721, D771, and R793, and L721 are 87, 93, and 94 % conserved, respectively). Figure 29A shows SDS-PAGE analysis of the purified proteins, all of which carry varying amounts of lower molecular weight contaminating proteins or degradation products.

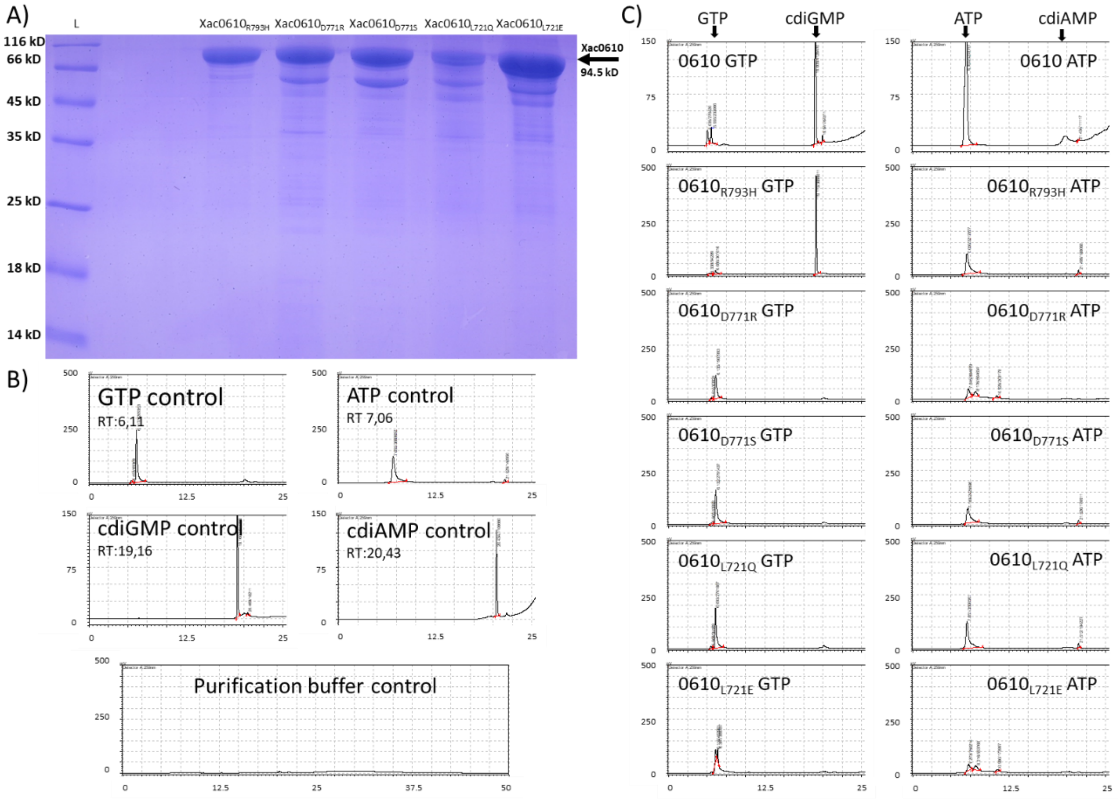


Figure 29. XAC0610 mutants are not able to use ATP as substrate. A) SDS-PAGE of 5 XAC0610 mutants used in enzymatic assays. B) HPLC of the substrates and expected products showing its retention time. Below, an HPLC ran only with XAC0610 showing absence of peaks. C) HPLC of the enzymatic reactions of each mutant. Column in the left shows the results using 1 mM GTP as a substrate and right column using 1 mM ATP as a substrate. Enzymatic reactions were done at 30 ºC with 2.5 uM ($Xac0610_{35\text{-}880}$ dimer) in a volume reaction of 100 ul containing 20 mM Tris-Cl pH 8.00, 5 mM MgCl2, 100 mM NaCl and 1 mM substrate. The reactions were quenched at 1 hour adding 900 ul of stop solution (10 mM HCl). The nucleotide products were analyzed by HPLC in the same form that in figure 10.

We performed enzymatic assays with $XAC0610_{35\text{-}880}$ and its 5 mutants using ATP or GTP as substrates. The presence of products was evaluated by HPLC in a C18 column (see methods). As seen in Figure 29B the retention times of GTP, ATP, cdiGMP and cdiAMP were 6.11, 7.06, 19.16 and 20.43 minutes respectively. When we tested the activity of XAC0610$_{35\text{-}880}$ against GTP or ATP we verified the specificity in the recognition of the guanine moiety (Fig. 29C).

Of the 5 mutants tested, XAC0610$_{35\text{-}880\_D771R}$, XAC0610$_{35\text{-}880\_D771S}$, XAC0610$_{35\text{-}880\_L721E}$, XAC0610$_{35\text{-}880\_L721Q}$ did not show activity with GTP or ATP. Only XAC0610$_{35\text{-}880\_R793H}$ retained the diguanylate cyclase activity but did not present diadenylate cyclase activity. It is also important to point out that although none of the mutants presented diadenylate cyclase activity, it is possible that one or more of the mutants can bind ATP in its active site. Other experiments are necessary to test this hypothesis.

1.4.13. Probable structural effects of the mutation in R793:

In an attempt to better understand why the XAC0610$_{R793H}$ mutant retained the diguanylate cyclase activity we produced structural models based on a model of GGDEF domain of XAC0610 produced by the Alpha fold software (Jumper et al. 2021). Figure 30A shows the hydrogen bonds formed by the side chain of R793 in the model of the wild-type XAC0610 protein.

The importance of this residue probably may lie in its involvement in the stabilization of the highly conserved GGDEF loop (AGDEF in XAC0610) through the interaction with the main chain carbonyl group of G796 and the side chain of the highly conserved D719 (Fig. 30A). Also, this residue could form a hydrogen bond with the C6 carbonyl group of the guanine moiety (as shown for the PleD model in complex with GTPαS, PDB ID: 2V0N). When this arginine is changed to histidine (Fig. 30B) we noted that this residue could form hydrogen bonds with the G796 and D719 residues, but may not be able to form a hydrogen bond with a carbonyl group of the guanine moiety.

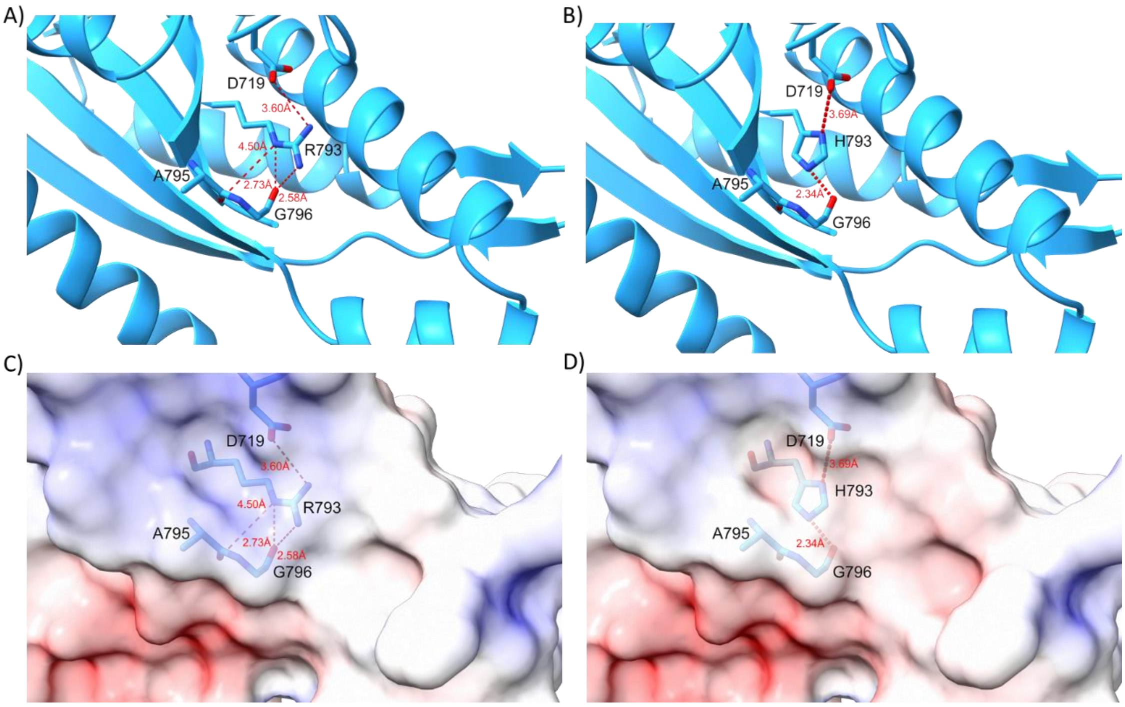
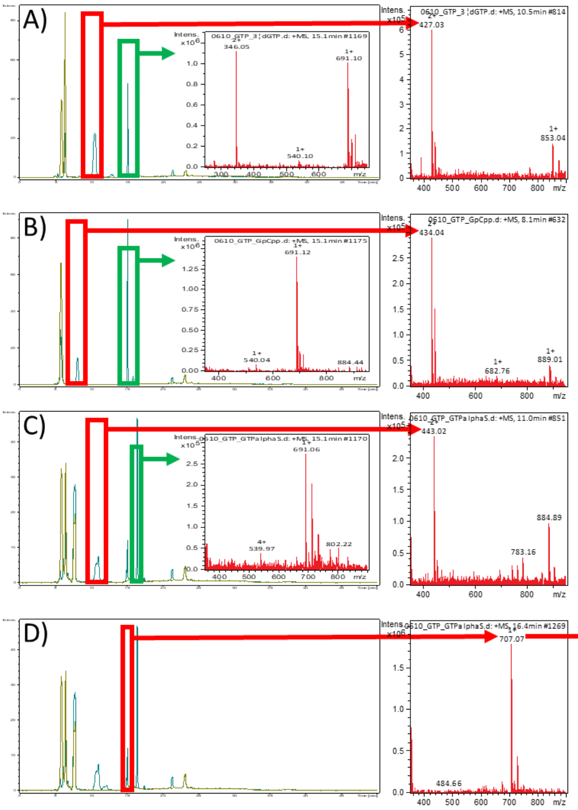


Figure 30. XAC0610$_{R793H}$ model to explain the diguanylate cyclase activity. A) Structural model of the guanine recognition site of XAC0610 showing the hydrogen bonds network of the R793 residue. B) Model of the R793H mutation and its hydrogen bonds network. Is interesting to note that hydrogens bonds are conserved between R793 and R793H. C) Electrostatic surface model of the same region in A) showing a positive region near to R793. D) Electrostatic surface model of the same region in B) showing a negative region near to R793H. Also, we can see a hole produced by the absence of the arginine.

Electrostatic surface models of XAC0610$_{wt}$ and XAC0610$_{R793H}$ show that in the wild-type protein the surface surrounding R793 has a more positive electrostatic potential than the surface surrounding R793H (Fig. 30C and 30D). We could expect that the more negative electrostatic potential surrounding the H793 residue could favour the binding of an adenine moiety. However, as we can see in the enzymatic assays, this solely mutation (R793H) did not allow XAC0610$_{35\text{-}880\_R793H}$ to catalyze the formation of cdiAMP from ATP.

1.4.14. XAC0610 Enzymatic assays using GTP analogs:



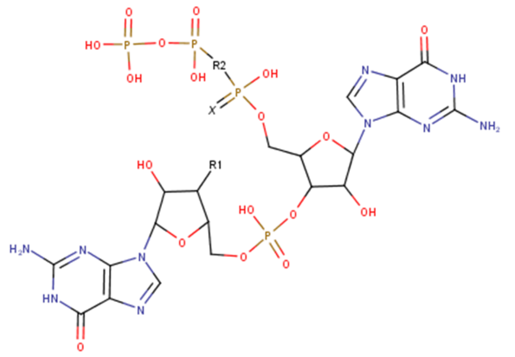


Fig.	Linear product	R1	R2	X	Molecular weight (g/mol)
8A	GTP-3´dGTP	-H	-O-	O	852.04
8B	GpCpp-GMP	-OH	$-CH_2-$	O	866.06
8C	GTPalphaS-GMP	-OH	-O-	S	884.02

c-GMP-GMPalphaS
706.07 g/mol

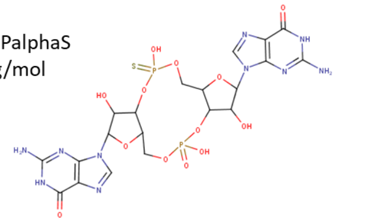


Figure 31. Enzymatic assays using GTP analogs. A) LC-MS analysis of the reaction with a 1:1 mix of GTP and 3´dGTP showing the GTP-3´dGMP peak formation as confirmed by mass spectrometry. B) LC-MS analysis of the reaction with a 1:1 mix of GTP and GpCpp showing the GpCpp-GMP peak formation as confirmed by mass spectrometry. C) LC-MS analysis of the reaction with a 1:1 mix of GTP and GTPalphaS showing the GTPalphaS-GMP peak formation as confirmed by mass spectrometry. D) LC-MS analysis of the reaction with a 1:1 mix of GTP and GTPalphaS showing the cGMP-GMPalphaS peak formation as confirmed by mass spectrometry. In all the HPLCs, the yellow chromatogram represents the substrates without protein and the green chromatogram represents the enzymatic reaction (protein plus substrates). Enzymatic reactions were done at 30 ºC with 2.5 µmol/L (Xac0610$_{35\text{-}880}$ dimer) in a volume reaction of 100 µL containing 20 mmol/L Tris-Cl pH 8.00, 5 mmol/L MgCl2, 100 mmol/L NaCl and 1 mmol/L substrate. The reactions were quenched at 16 hours adding 900 µL of stop solution (10 mmol/L HCl). HPLC and reaction conditions were the same as figure 10.

In order to gain new insights about the enzymatic mechanism of cdiGMP formation we performed enzymatic assays using a 1:1 mix of GTP and GTP analogs (3'dGTP, GpCpp and GTPalphaS). We observed the formation of three linear products GTP-3´dGMP (Fig. 24A), ppCpGpG (Fig. 31B), GTPalphaS-GMP (Fig. 31C) and a cyclic product cGMP-GMPalphaS (Fig. 31D). The formation of the cyclic product (cGMP-GMPalphaS) shows that oxygen in the position *pro-R* or *pro-S* is not completely necessary for the catalytic activity. We need to use stereochemically pure GTPalphaS *pro-R* or GTPalphaS *pro-S* as substrates to understand which of these positions (if any) is really important for the enzymatic reaction.

1.4.15. XAC0610 Enzymatic assays using 2´dGTP:

In addition to the enzymatic assays using ATP or GTP, we also tested if 2'dGTP could act as a substrate for XAC0610. These enzymatic reactions were done with 2.5 µmol/L (XAC061035-880 dimer) in a volume reaction of 100 ul containing 20 mmol/L Tris-Cl pH 8.00, 5 mmol/L MgCl2, 100 mmol/L NaCl and 1 mmol/L substrate. The reactions were quenched at 16 hours adding 900 ul of stop solution (10 mmol/L HCl). Interestingly, we verified that 2'dGTP acts as a substrate producing cdi2'dGMP (Fig. 32B) as shown by the mass spectrometry profile (Fig. 32B). When we carried out a reaction using a 1:1 mix of GTP and 2'dGTP we observed the production of a cyclic hybrid molecule c-GMP2'dGMP (Fig. 32C).

A deeper analysis in the peak of this product shows that in fact when we mix the two substrates, three cyclic products were produced (cdiGMP, cGMP2'dGMP, cdi2'dGMP) (Fig. 32D). Interestingly, when reactions were carried out for shorter times (10 minutes), XAC0610$_{35\text{-}880}$ produced two main products: cdiGMP and the linear dinucleotide pppGp2´dG or ppp2´dGpG (Fig. 33).

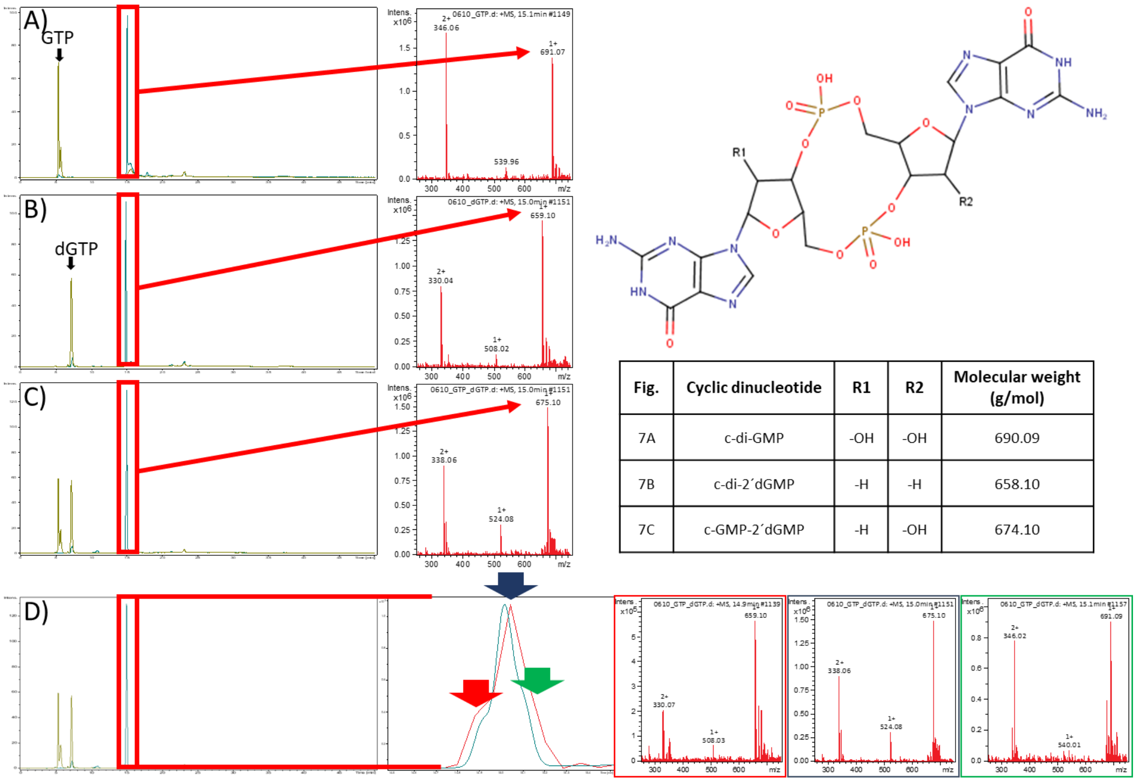
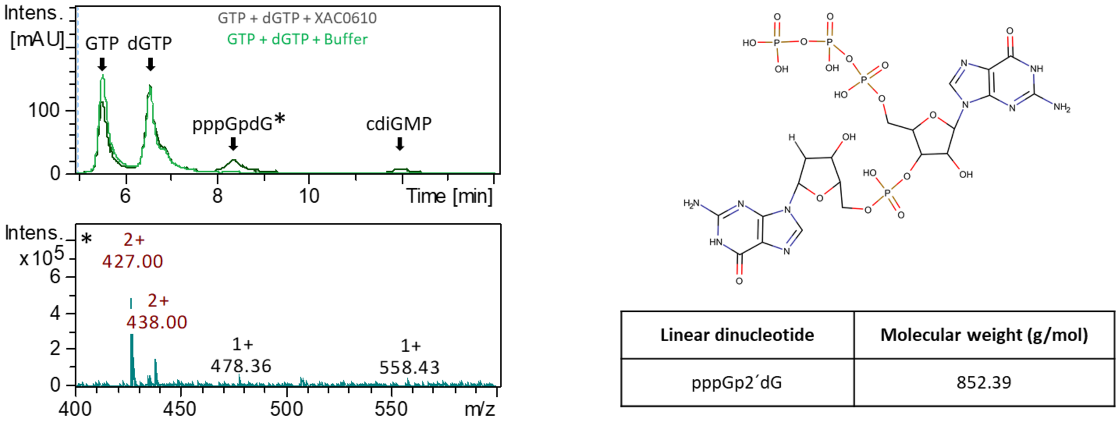


Fig.	Cyclic dinucleotide	R1	R2	Molecular weight (g/mol)
7A	c-di-GMP	-OH	-OH	690.09
7B	c-di-2´dGMP	-H	-H	658.10
7C	c-GMP-2´dGMP	-H	-OH	674.10

Figure 32. Enzymatic assays using 2´dGTP. A) LC-MS analysis of the reaction with GTP showing the cdiGMP peak formation as confirmed by mass spectrometry. B) LC-MS analysis of the reaction with 2´dGTP showing the cdi2´dGMP peak formation as confirmed by mass spectrometry. C) LC-MS analysis of the reaction with a 1:1 mix of GTP and 2´dGTP showing the cGMP-2´dGMP peak formation as confirmed by mass spectrometry. D) A zoom in the peak of products of the HPLC chromatogram showed in C showing that the peak is in fact formed by three superposed peaks. As shown in the right, analysis of each region (red, blue or green) shows different mass spectrometry profiles corresponding to the three possible products. Enzymatic reactions were done as in Fig. 31. In all the HPLCs, the yellow chromatogram represents the substrates without protein and the green chromatogram represents the enzymatic reaction (protein plus substrates).

It is more likely that the linear product formed was pppGp2´dG due to the absence of the 2´-OH in the 2´dGTP that could reduce the nucleophilicity of the 3´-OH of 2´dGTP. Thus, even though the enzyme binds the two substrates in different monomers, probably the attack of 3´-OH of dGTP to the α-phosphate of GTP occurs more slowly than attack of 3´-OH of GTP to α-phosphate of dGTP.



Linear dinucleotide	Molecular weight (g/mol)
pppGp2´dG	852.39

Figure 33. 10 minutes XAC0610 enzymatic reactions using GTP and 2´dGTP. A) LC-MS analysis of the reaction with a 1:1 mix of GTP and 2´dGTP in a 10 minutes reaction showing the accumulation of pppGp2´dG as confirmed by mass spectrometry (below). B) A cartoon of the chemical configuration of pppGp2´dG showing the theoretical molecular weight. Enzymatic reactions were done as in Fig. 31

To our knowledge this is the first description of the enzymatic production of cGMP2'dGMP and pppGp2´dG. Since 2'dGTP is a natural molecule found in all cells, this raises the possibility that these product (as well cdi2'dGMP) could be formed *in vivo* and could represent new signaling molecules of the cyclic dinucleotides group.

1.4.16. Kinetic studies of XAC0610 with the substrates 2´dGTP and GTP:

To gain insight on the kinetics of these reactions, we first formulated a simple kinetic model of GGDEF proteins in the presence of GTP or dGTP (Fig. 34A). Both of them have two dissociation constants, one for the binding of the first substrate (K_{G1} or K_{dG1}) and other for the binding of the second substrate (K_{G2} or K_{dG2}). Furthermore, each model has a K_{cat} corresponding to the formation of its respective products. When GTP is present as a substrate, only cdiGMP product is formed (K_{cat_cdiG}). On the other hand, we have no information of which products are formed in short reactions when just dGTP is present as a substrate. For this reason, we named the K_{cat} in presence of just dGTP as K_{cat_dG} meaning that the product could be a linear (pppdGpdG) or a cyclic product (cdidGMP). Additionally, to avoid confusions, all the initial velocities (Vo) and K_{cat} are expressed in number of PPi released per enzyme dimer per second (equivalent to number of nucleophilic attacks per enzyme dimer per second).

To estimate the posterior distributions of the dissociation constants for GTP (K_{G1} and K_{G2}) and dGTP (K_{dG1} and K_{dG2}) and the turnover numbers k_{cat_cdiG} and k_{cat_dG}, we performed enzymatic assays using initial different GTP or dGTP concentrations and measured PPi release using an enzyme-coupled spectrophotometric method. Using Bayesian inference and MCMC methods we fitted the experimental results to a model for a dimer protein described by (Oliveira et al., 2015) with GTP or dGTP as substrates (Fig. 34, 35). Uninformative prior distributions were used for these analyses (Table 6).

When GTP was used as a substrate, we obtained a k_{cat_cdiG} with mean 0.2741 s^{-1} (0.2715 – 0.2765 High Posterior Densiity (HPD) 95 %), a K_{G1} of 130.3984 µM (44.4549 – 196.8094 HPD 95 %) and a K_{G2} of 1.5193 µM (0.8375 – 3.1855 HPD 95 %) (Fig. 26, 27, Table 7). Enzymatic assays with just 2´dGTP as substrate were difficult to perform due the small amounts of PPi detected. The mean of k_{cat_dG} was 0.0028 s^{-1} (0.0023 – 0.0034 HPD 95 %). K_{dG1} and K_{dG2} had a mean of 61.3160 µM (1.0127 – 188.0275 HPD 95 %) and 0.3689 µM (0.0108 – 1.7862 HPD 95 %), respectively (Fig. 34, 35, Table 7).

A

$$E + G \underset{}{\overset{K_{G1}}{\rightleftarrows}} E_G + G \overset{K_{G2}}{\rightleftarrows} E_G^G \xrightarrow{Kcat_{cdiG}} E + cdiG$$

$K_{cat\_cdiG}\ (s^{-1}) = 0.2583\ - 0.2925$

$K_{G1}\ (\mu M) = 44.4549\ - 196.8094$

$K_{G2}\ (\mu M) = 0.8375\ - 3.1855$

$$E\ + dG \overset{K_{dG1}}{\rightleftarrows} E_{dG} + dG \overset{K_{dG2}}{\rightleftarrows} E_{dG}^{dG} \xrightarrow{Kcat_{dG}^*} E + cdidG$$

$K_{cat\_dG}\ (s^{-1})^* = 0.0023\ - 0.0034$

$K_{dG1}\ (\mu M) = 1.0127\ - 188.0275$

$K_{dG2}\ (\mu M) = 0.0108\ - 1.7863$

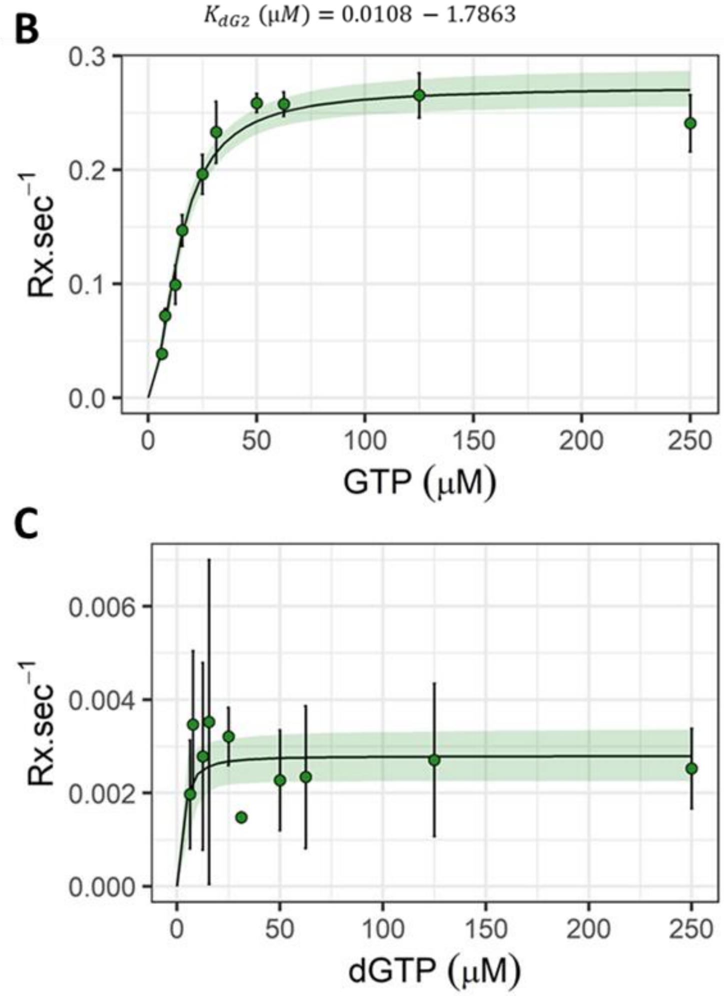


Figure 34. Kinetic model of a homodimer with one substrate. A) Above, model when GTP is used as a substrate. Below, model when 2´dGTP is used as the substrate. High posterior densities 95 % of their parameters are showed below the models. B) Results of the Bayesian inference fitting of initial velocity vs. GTP concentration. C) Results of the Bayesian inference fitting of initial velocity vs. dGTP concentration. In B and C points reflects the mean of three independent experiments with bars showing their respective standard deviations. Line and ribbons correspond to the mean and 95 % HPD intervals of the fitting.

Table 6. Prior distributions for the estimations of all the model parameters.

Prior Distributions			
Parameter	**Distribution**	**LB**	**UB**
Cooperative homodimer model			
GTP			
K_{cat_cdiG} (s^{-1})	Uniform	0	1
K_{G1} (μM)	Uniform	0	200
K_{G2} (μM)	Uniform	0	200
dGTP			
K_{cat_dG} (s^{-1})*	Uniform	0	0,3
K_{dG1} (μM)	Uniform	0	200
K_{dG2} (μM)	Uniform	0	200
GTP - dGTP			
K_{cat_cdiG} (s^{-1})	Uniform	0,2583	0,2925
$K_{cat\_pppGpdG}$ (s^{-1})	Uniform	0	1
K_1 (μM)	Uniform	0	200
K_2 (μM)	Uniform	0	200
K_4 (μM)	Uniform	0	200
K_5 (μM)	Uniform	0	200
K_6 (μM)	Uniform	0	200
Non-cooperative homodimer model			
GTP			
K_{cat_cdiG} (s^{-1})	Uniform	0	1
K_G (μM)	Uniform	0	200
dGTP			
K_{cat_dG} (s^{-1})*	Uniform	0	0,3
K_{dG} (μM)	Uniform	0	200
GTP - dGTP			
K_{cat_cdiG} (s^{-1})	Uniform	0,2714	0,3319
$K_{cat\_pppGpdG}$ (s^{-1})	Uniform	0	1
K_G (μM)	Uniform	0	200
K_{dG} (μM)	Uniform	0	200

*k_{cat_dG} denotes that the product can be the linear or the cyclic product

Table 7. Summary of posterior distributions of parameters of one substrate model.

Cooperative homodimer model (one substrate)			
Parameter	**Mean**	**L-95 % CI**	**U-95 % CI**
K_{cat_cdiG} (s^{-1})	0.2741	0.2583	0.2925
K_{G1} (µM)	130.3984	44.4549	196.8094
K_{G2} (µM)	1.5193	0.8375	3.1855
K_{cat_dG} (s^{-1})*	0.0028	0.0023	0.0034
K_{dG1} (µM)	61.3160	1.0127	188.0275
K_{dG2} (µM)	0.3689	0.0108	1.7862

*k_{cat_dG} denotes that the product can be the linear or the cyclic product

Taken together, K_{G1} was estimated to be at least 10 times higher than K_{G2}, k_{cat_dG} was approximately 100 times less than k_{cat_cdiG}. Due to the the low value of k_{cat_dG}, measurement errors with low concentrations of dGTP were very large (Fig. 34), resulting in an uncertain posterior distribution of K_{dG1} and K_{dG2} with the results being that K_{dG1} could be equal, much greater or even less than K_{dG2}.

After determining kinetic constants for homodimer models with one substrate type (GTP or dGTP), and because we demonstrate that XAC0610 can form hybrid products (cGMPdGMP or pppGpdG), we proposed a homodimer model with two substrates (GTP and dGTP) (Fig 36A). Additionally, the above results suggest that dGTP retains the ability to act as an electrophile at its α-phosphate but has essentially lost its ability to act as a nucleophile at the 3´-OH group. Thus, we can simplify the model removing from it the formation of cdidGMP and assuming that when one GTP binds to one monomer and dGTP to the other monomer, only the linear product pppGpdG is formed (Fig 36B). If this is correct, then in reactions where both GTP and dGTP are simultaneously present, XAC0610 will form just pppGpdG and/or cdiGMP.

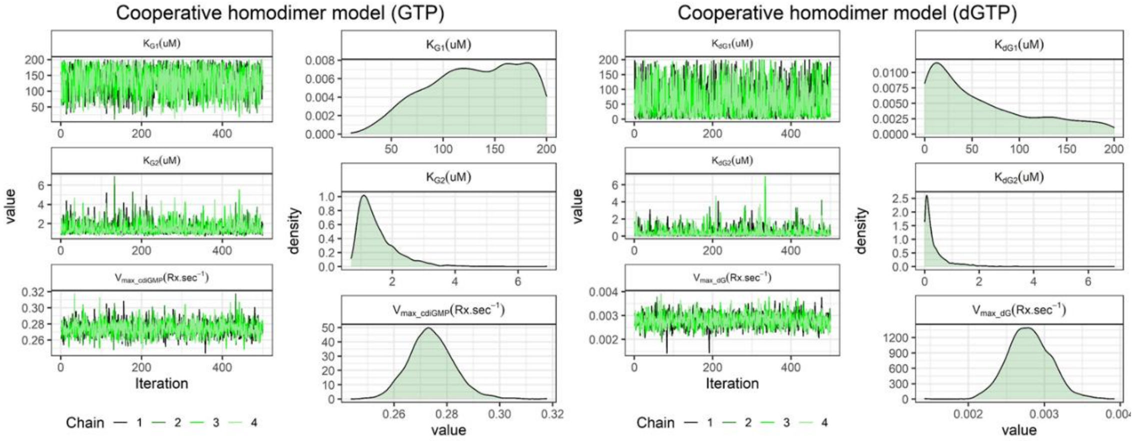
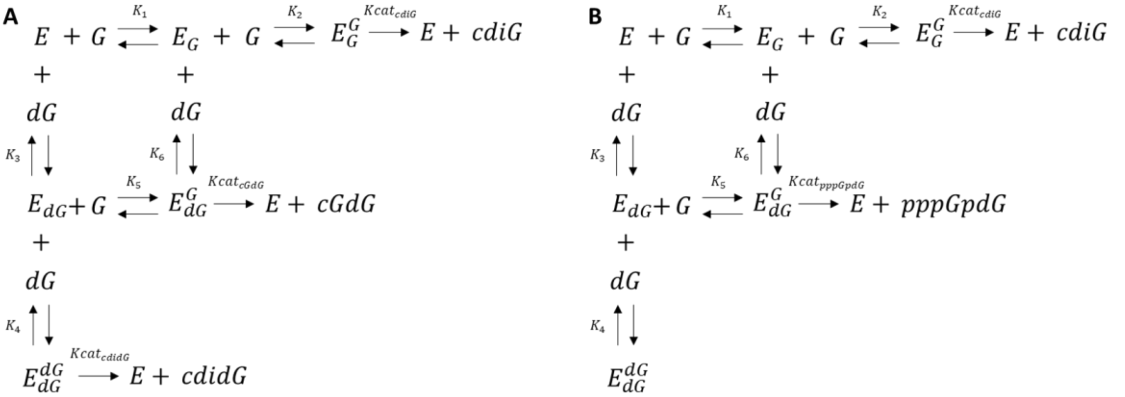


Figure 35. Markov Chain Monte Carlo (MCMC) chains and posterior densities of paramaters of the model in figure 26A. Left columns show the four MCMC chains of 500 iterations for each parameter. Right column is showing the posterior density of each parameter.


A
B
$E + G \rightleftharpoons^{K_1} E_G + G \rightleftharpoons^{K_2} E_G^G \xrightarrow{Kcat_{cdiG}} E + cdiG$
K_3
K_6
K_5
$E_{dG} + G \rightleftharpoons E_{dG}^G \xrightarrow{Kcat_{cGdG}} E + cGdG$
K_4
$E_{dG}^{dG} \xrightarrow{Kcat_{cdidG}} E + cdidG$
$E_{dG} + G \rightleftharpoons E_{dG}^G \xrightarrow{Kcat_{pppGpdG}} E + pppGpdG$
E_{dG}^{dG}

$$V_o total = \frac{V_{\max(cdiG)}K_4K_6[G]^2 + V_{\max(cGdG)}K_2K_4[G][dG] + V_{\max(cdidG)}K_2K_5[dG]^2}{K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}$$

$$V_o total = \frac{V_{\max(cdiG)}K_4K_6[G]^2 + V_{\max(pppGpdG)}K_2K_4[G][dG]}{K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}$$

Figure 36. Homodimer model with two substrates and its respective equations. A) Model of homodimer enzymes with two possible susbtrates (GTP (G) and dGTP (dG)). Three possible cyclic products are considered. B) Model in B simplified due to the low Kcat_dG observed. Just two products are allowed in this model (cdiGMP and pppGpdG) that results for the nucleophilic attack of GTP to other GTP or to dGTP. Equations derived from these models assuming steady state and preequilibrium are showed below the model.

To determine the fraction of pppGpdG and cdiGMP that is produced, we performed one-minute reactions using different initial GTP concentrations (31.25 – 250 µM) and three fixed initial dGTP concentrations (0, 100, 200, and 300 µM). The products of these reactions were analyzed by HPLC-ESI-MS (Fig. 37A). These assays showed that, at fixed dGTP concentrations, the cdiGMP fraction increases linearly whereas the pppGpdG fraction decreases linearly with respect to the GTP concentration (Fig. 37B). Additionally, we observed that when initial concentrations of the two substrates are similar, the products (pppGpdG and cdiGMP) are formed in at similar concentrations (Fig. 37B).

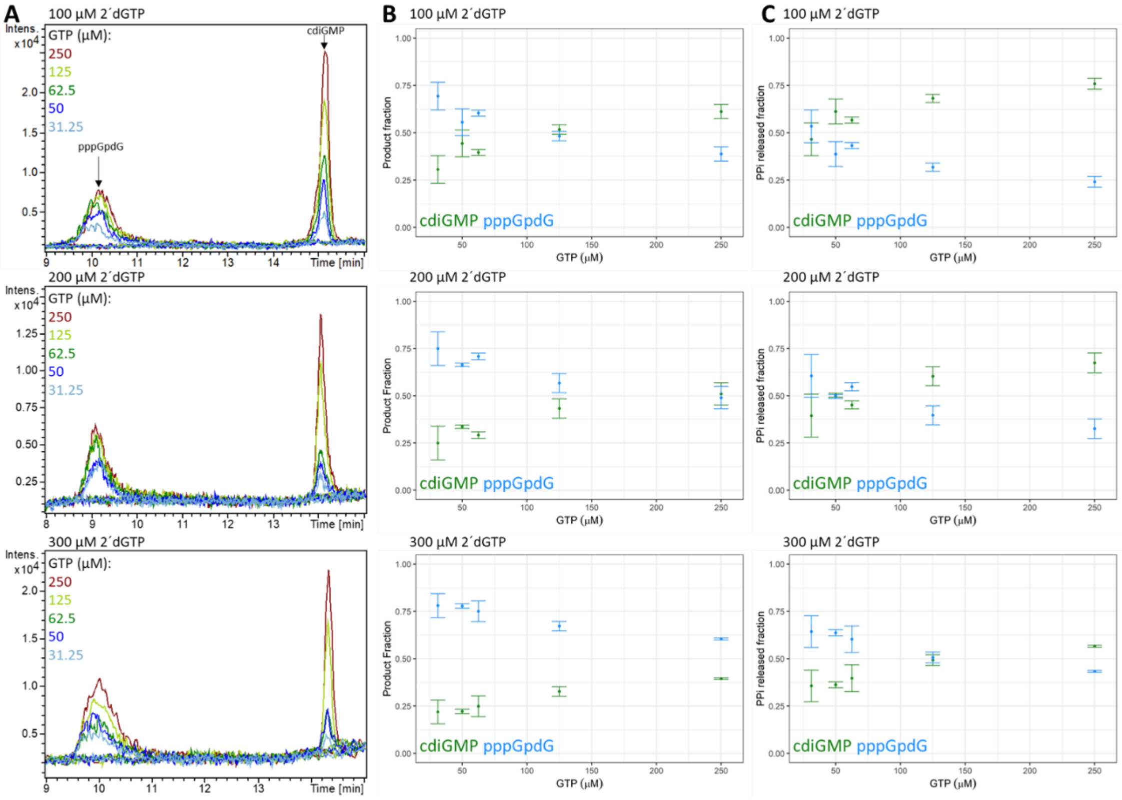


Figure 37. Analysis of ratios of pppGpdG and cdiGMP in reactions where GTP and dGTP are present in different initial concentrations. A) Extracted ion chromatograms of the peaks corresponding to the m/z ratio for cdiGMP and pppGpdG. B) Using the peak areas of A, we calculated the fractions of cdiGMP and pppGpdG formed. C) Multiplying by two the area corresponding to cdiGMP, we estimated the fractions of released PPi corresponding to the formation of pppGpdG and cdiGMP. First, second and third rows are showing reactions with 100, 200, and 300 µM initial dGTP, respectively. One-minute reactions were performed to obtain the peaks in A.

To explain how the linear product pppGpdG and the cyclic product cdiGMP is formed when GTP and 2´dGTP are present, we performed enzymatic assays measuring the PPi release in reactions with different initial GTP concentrations (0 – 250 µM) in the presence of three fixed concentrations of dGTP (100, 200, and 300 µM). Using standard steady-state and preequilibrium assumptions, we derived the rate equation for the PPi production as a function of GTP and 2´dGTP (Fig. 37B, Appendix 1.1). In this equation each nucleophilic attack (and its corresponding PPi release) represents one reaction. Thus, the production of cdiGMP from two GTP is produced by two reactions (two PPi released), whereas the production of pppGpdG from GTP and 2´dGTP is produced by just one reaction (one PPi released).

Using the calculated initial velocity (Vo) from the enzymatic assays in different dGTP and GTP concentrations (Vo_{total}) (Fig. 38C) and the fraction of PPi released by the formation of pppGpdG or cdiGMP from the experiments (Fig. 37C), we calculated the initial velocity of formation of pppGpdG ($Vo_{pppGpdG}$) and cdiGMP (Vo_{cdiGMP}) (Fig. 37C). These calculations were done by multiplying the fraction of PPi released by each product with the Vo_{total} in the same dGTP and GTP concentrations.

To estimate the posterior distributions of the parameters that explain the homodimer model with two substrates (Fig. 36), we performed a multivariate Bayesian inference (three response variables: Vo_{total}, Vo_{cdiGMP}, and $Vo_{pppGpdG}$) and MCMC (Fig. 38). We used uninformative priors for all the parameters except for the K_{cat_cdiG} where the prior distribution was adjusted to the results obtained when only GTP was used as a substrate

(Table 6). The posterior distributions of the eight constants that explain the model in Figure 36B (K_1, K_2, K_3, K_4, K_5, K_6, K_{cat_cdiG} and $K_{cat\_pppGpdG}$) are summarized in Table 8.

Table 8. Summary of posterior distributions of parameters of two substrate model.

Cooperative homodimer model (two substrates)			
Parameter	**Mean**	**L-95 % CI**	**U-95 % CI**
K_{cat_cdiG} (s^{-1})	0,2664	0,2596	0,2742
$K_{cat\_pppGpdG}$ (s^{-1})	0,2173	0,2025	0,2339
K_1 (µM)	145,4261	71,7880	196,5760
K_2 (µM)	1,3397	0,9049	2,3833
K_3 (µM)*	118,6835	26,1993	490,7323
K_4 (µM)	5,1538	0,6817	13,7367
K_5 (µM)	2,0224	0,2852	5,2659
K_6 (µM)	1,0041	0,6750	1,7782

Our findings indicate that we have a well fitted theoretical model (Fig. 39) in which $K_{cat\_pppGpdG}$ is slightly slower than K_{cat_cdiG} (mean of 0.2173 vs. 0.2664 s^{-1}). The formation of the linear product is explained due to a change in affinities for GTP and 2´dGTP of XAC0610 depending on the state of the monomers. When the two XAC0610 GGDEF domains are empty, the affinities for GTP and dGTP are very similar (mean of 145.4261 µM for K_1 vs. 118.6835 µM for K_3, respectively). Similarly, the dissociation constants for GTP and dGTP are very similar when one monomer has already bound GTP (mean of 1.3397 µM for K_2 vs. 1.0041 µM for K_6, respectively). However, when one monomer has bound dGTP, our model showed that the empty monomer has a two and a half-fold greater affinity for GTP than for dGTP (mean of 2.0224 µM for K_5 vs. 5.1538 µM for K_4, respectively).

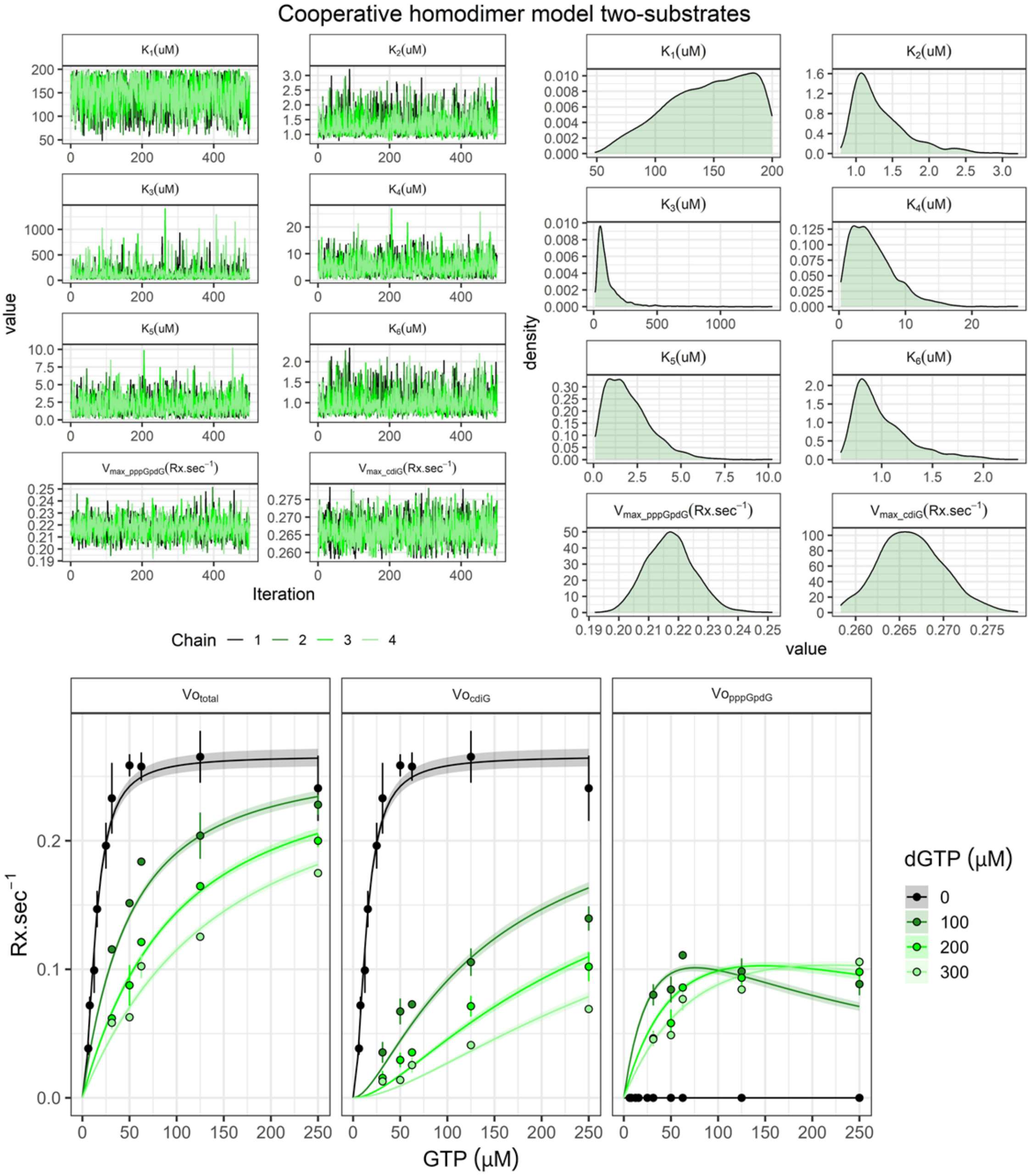
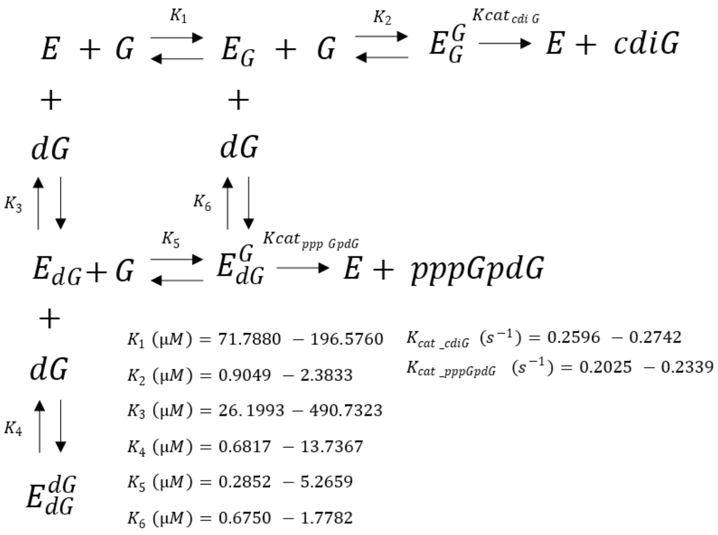


Figure 38. Two substrates homodimer model bayesian fitting. A) MCMC chains for the parameter estimation of the model. B) Posterior distributions of each of the parameters related to the two substrates homodimer model. C) Posterior distributions of the parameters were used to fit the experimental data to the proposed model. We estimate the values of the three types of velocities.


K_1
K_2
$Kcat_{cdi\,G}$
$E + G \rightleftarrows E_G + G \rightleftarrows E_G^G \longrightarrow E + cdiG$
$+$
dG
K_3
K_6
K_5
$Kcat_{ppp\,GpdG}$
$E_{dG} + G \rightleftarrows E_{dG}^G \longrightarrow E + pppGpdG$
$+$
dG
K_4
E_{dG}^{dG}

$K_1\ (\mu M) = 71.7880 - 196.5760$

$K_2\ (\mu M) = 0.9049 - 2.3833$

$K_3\ (\mu M) = 26.1993 - 490.7323$

$K_4\ (\mu M) = 0.6817 - 13.7367$

$K_5\ (\mu M) = 0.2852 - 5.2659$

$K_6\ (\mu M) = 0.6750 - 1.7782$

$K_{cat\_cdiG}\ (s^{-1}) = 0.2596 - 0.2742$

$K_{cat\_pppGpdG}\ (s^{-1}) = 0.2025 - 0.2339$

Figure 39. GTP - dGTP homodimer model of XAC0610. Enzymatic model that explains the formation of cdiGMP and pppGpdG when XAC0610 (E) is present in presence of GTP (G) and dGTP (dG). 95 % HPD intervals are shown for each parameter.

Taken together, our model proposes that the affinity for GTP or dGTP of XAC0610 varies depending on if the GGDEF dimer is in a free state (two monomers empty) or which substrate (GTP or dGTP) is bound to one of the monomers. In the free state or if one of its monomers have GTP, XAC0610 apparently could bind any of the substrates (GTP or dGTP). But if one of the monomers has bound dGTP, the empty monomer prefers GTP resulting in the observed fractions of pppGpdG and cdiGMP when the two substrates are present.

We further explored if we could explain the experimental data using a non-cooperative model (K1 = K2 = K5 = KG, and K3 = K4 = K6 = KdG). Fitting using the same methods used for the cooperative model results in similar distributions for kcat_cdiG and kcat_pppGpdG and similar distributions for KG and KdG (with a slightly greater dissociation constant for KdG) (Fig. 40-42, Table 9-10).

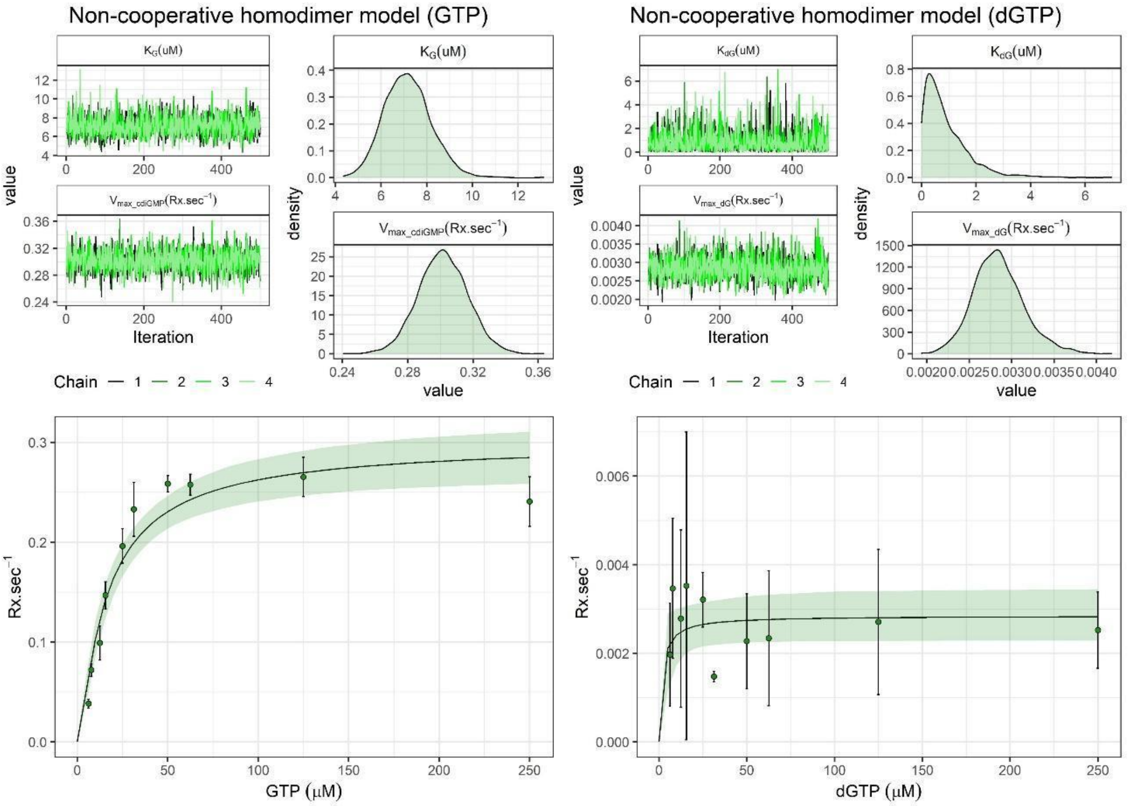


Figure 40. Non-cooperative onse substrate model fitting. MCMC chains, posterior densities and fitting results of the Bayesian inference assuming no cooperativity for the models presented in figure 34A.

Table 9. Summary of posterior distributions of parameters of non-cooperative one substrate model.

Non-cooperative homodimer model (one substrate)			
Parameter	**Mean**	**L-95 % CI**	**U-95 % CI**
K_{cat_cdiG} (s^{-1})	0,3016	0,2714	0,3319
K_G (μM)	7,1618	5,3404	9,2347
K_{cat_dG} (s^{-1})	0,0028	0,0023	0,0035
K_{dG} (μM)	0,9057	0,0416	3,2324

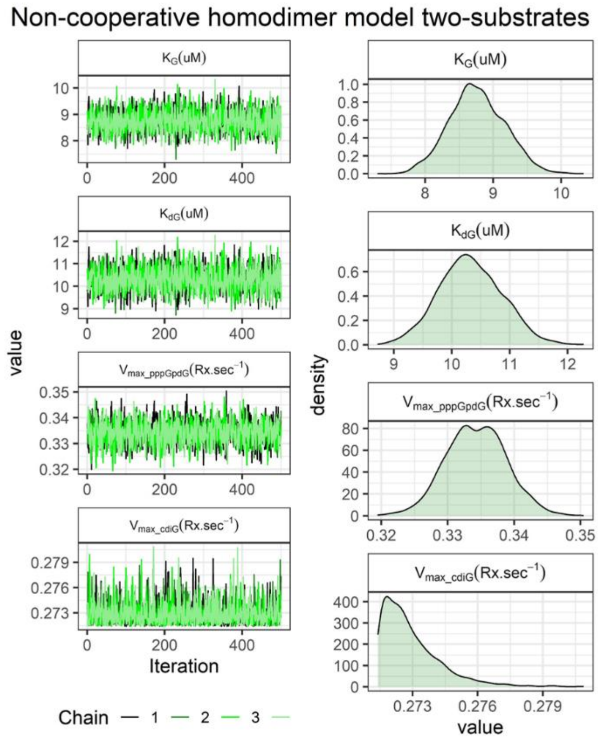


Figure 41. MCMC chains and posterior densities of paramaters of the non-cooperative homodimer model with two substrates. Left column shows the MCMC chains of 500 iterations for each parameter. Right column is showing the posterior density of each parameter.

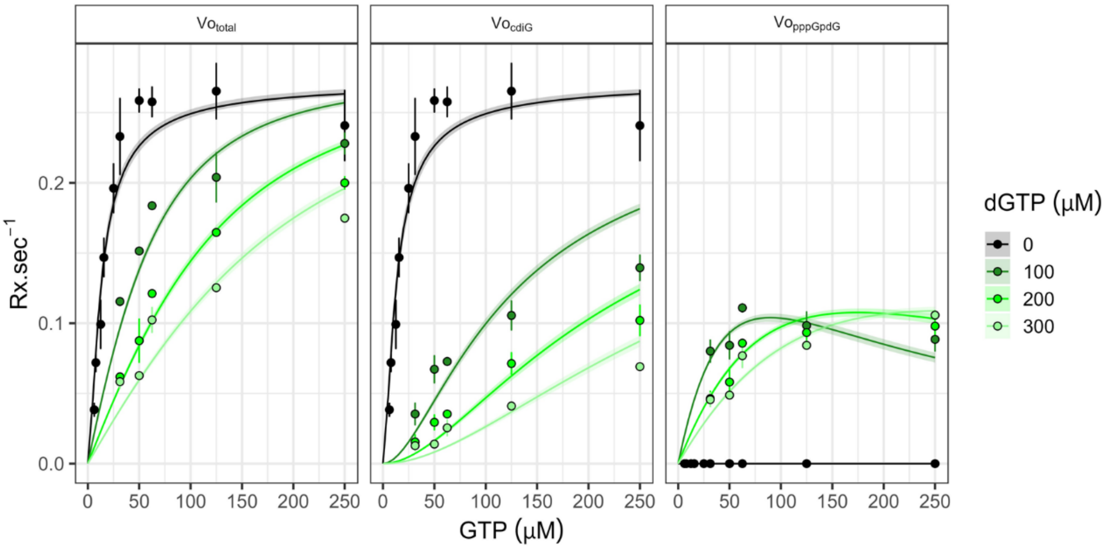


Figure 42. Non-cooperative homodimer model with two substrates fitting. Posterior distributions of the parameters were used to fit the experimental data to the proposed model. We estimated the values of the three types of velocities.

Table 10. Summary of posterior distributions of parameters of non-cooperative two substrate model.

Non-cooperative homodimer model (two substrates)			
Parameter	**Mean**	**L-95 % CI**	**U-95 % CI**
K_{cat_cdiG} (s^{-1})	0,2729	0,2715	0,2765
$K_{cat\_pppGpdG}$ (s^{-1})	0,3343	0,3254	0,3431
K_G (μM)	8,7631	7,9672	9,5735
K_{dG} (μM)	10,3239	9,2911	11,3884

To determine which of the two models (cooperative or non-cooperative) predict better the data, we estimated the expected log pointwise predictive density (elpd) using leave one-out (loo) cross-validation (Vehtari et al., 2017). The results showed a significant difference (more than two times the standard error) in favor of the cooperative model (Table 11).

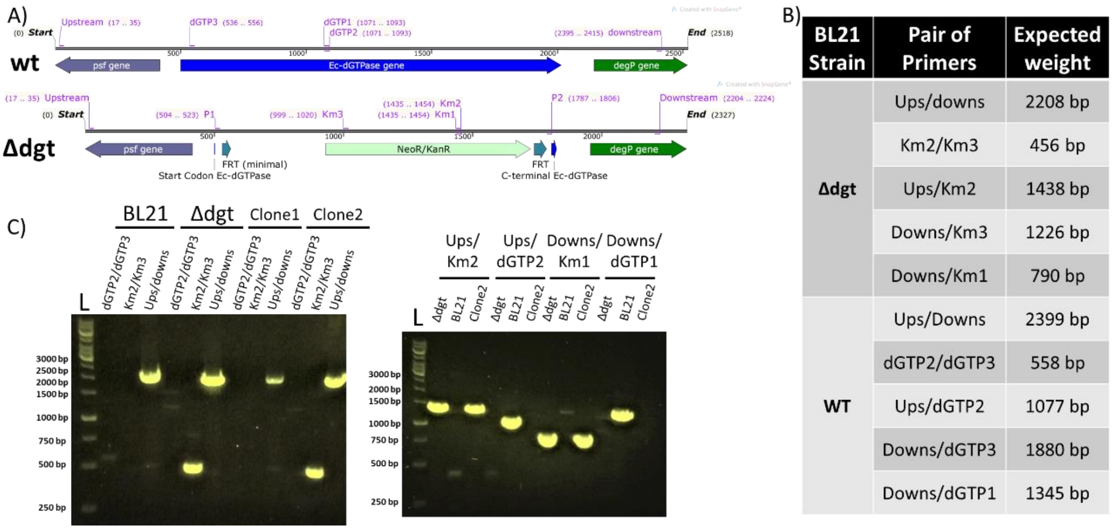
Table 11. Cooperative and non-cooperative model comparison.

Model Comparison		
Model	**elpd_diff**	**se_diff**
Homodimer one substrate (GTP)		
Cooperative	0,0	0,0
Non-cooperative	-7,1	1,3
Homodimer one substrate (dGTP)		
Cooperative	-0,2	0,2
Non-cooperative	0,0	0,0
Homodimer two substrates (GTP - dGTP)		
Cooperative	0,0	0,0
Non-cooperative	-37,7	3,1

1.4.17. Testing for pppGp2´dG production *in vivo*:

The above analysis shows that in presence of similar amounts of GTP and 2´dGTP, similar amounts of cdiGMP and pppGpdG are formed. *In vivo* levels of GTP are generally significantly greater than 2´dGTP (Buckstein et al., 2008; Ferraro et al., 2009; Traut, 1994). However, the ratio of GTP and 2´dGTP can change in several sitations. For instance, intracellular levels of 2´deoxynucleotides can be controlled to avoid mutational events. For example, in *E. coli*, the enzyme *Ec*-dGTPase degrades excess 2´dGTP in certain metabolic situations (Barnes et al., 2019; Seto et al., 1988). Furthermore, some phages such as bacteriophage T7 have developed mechanisms to inhibit these regulation mechanisms to increase 2´deoxynucleotide pools, thereby allowing more efficient viral replication (Buckstein et al., 2008; Huber et al., 1988). Bacteriophage T7 codes for a peptide that inhibits *Ec*-dGTPase and at the same time increases the affinity for GTP, although this complex does not hydrolyze GTP (Nakai & Richardson, 1990). Thus, increased concentration of 2´dGTP caused by phage infection could lead to certain GGDEF proteins recognizing and employing 2´dGTP as a substrate. While the *X. citri* genome does not code for Ec-dGTP, pppGp2´dG production by functional homologs or analogs of XAC0610 could in this way function as a signaling molecule of viral infection in other bacterial species.

To test this hypothesis, we knocked out the gene coding for *Ec*-dGTPase in *E. coli* BL21(DE3), producing the ΔdGTPase (Δ*dgt*) strain (Fig. 43A). We verified the mutation of the *dGTPase* region using different sets of primers in the wild-type and the Δdgt strain (Fig. 43B-D). We then transformed wild-type and Δ*dgt* with pET28(a)-XAC0610$_{35\text{-}880}$ in order to detect which cyclic or linear dinucleotides were produced under inducing conditions. Total dinucleotide extraction from the wild-type and Δ*dgt* strains overexpressing XAC0610$_{35\text{-}880}$ was performed and mass-spectrometry was used to detect the presence of molecular ions with the expected mass-to-charge ratio (m/z) of cdiGMP (346^{+2} or 691^{+1}) and pppGp2´dG (427^{+2},438^{+2} or 853^{+1}) (Fig. 44B).



BL21 Strain	Pair of Primers	Expected weight
Δdgt	Ups/downs	2208 bp
	Km2/Km3	456 bp
	Ups/Km2	1438 bp
	Downs/Km3	1226 bp
	Downs/Km1	790 bp
WT	Ups/Downs	2399 bp
	dGTP2/dGTP3	558 bp
	Ups/dGTP2	1077 bp
	Downs/dGTP3	1880 bp
	Downs/dGTP1	1345 bp

Figure 43. Construction of a *E. coli Ec*-dGTPase Knock-out strain (Δdgt). A) Schematic representation of the genomic context of Ec-dGTPase of the BL21(DE3) wild-type (above) or Δdgt (below). Genes are indicated as arrows and primer binding sites are denoted with purple thin lines and their respective names. B) Primers used to confirm the production of the BL21(DE3) Δdgt strain. Right column shows the expected weight (in number of base pairs) of the PCR amplicon when using the corresponding pair of primers (middle column) with the corresponding strain (left column). C and D) Agarose-gel electrophoresis showing the expected patterns of bands to confirm that clone 2 corresponds to the BL21(DE3) Δdgt strain. Gel on the left is showing results from different set of primers (vertical letters in each band) using different strains (horizontal letters). From left to the right the strains are BL21(DE3) wild-type (BL21), *E. coli* donor strain Δdgt (Δdgt), BL21(DE3) that did not present the Ec-dGTPase mutation (clone 1) and BL21(DE3) Δdgt obtained by us (clone 2). Gel on the right also shows results from different set of primers (horizontal letters above the line) with different strains (Vertical letters below the line). Names of the strains are the same as on the gel on the left.

In both strains, induction of XAC0160 was confirmed by SDS-PAGE (Fig. 44A). Mass spectrometry analysis of the cellular extracts detected peaks corresponding to the expected weights of both cdiGMP and pppGp2´dG (Fig. 44B). However, peaks apparently belonging to pppGp2´dG in the *Δdgt* strain were repetitively less than for the wild-type strain (Fig. 44B). Ions corresponding to the m/z ratio of pppGp2´dG were detected, however, the number of ions with this m/z ratio was very low and we could not detect the fragmentation pattern to confirm that these ions corresponded to pppGpdG.

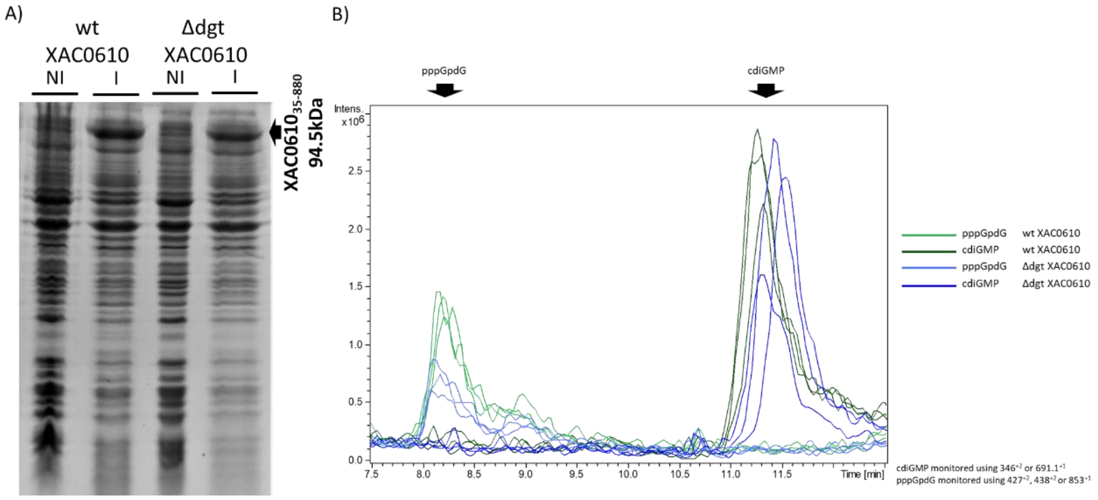


Figure 44. Analysis of the *in vivo* production of pppGp2´dG by BL21(DE3) wild-type and Δdgt strains overexpressing XAC0610$_{35\text{-}880}$. A) SDS-PAGE gel showing the induction of XAC0610$_{35\text{-}880}$ in the BL21(DE3) wild-type and Δdgt. A protein band corresponding to XAC0610$_{35\text{-}880}$ is visible after IPTG addition (I) but not before IPTG addition (NI). B) Ion chromatograms of the cyclic dinucleotides extractions from BL21(DE3) wild-type and Δdgt strains overexpressing XAC0610$_{35\text{-}880}$ (green and blue, respectively). The ion chromatograms are showing the signal of ions with m/z that match to the theoretical m/z of cdiGMP (346^{+2} or 691^{+1}) (dark green and blue green) and pppGp2´dG (427^{+2}, 438^{+2} or 853^{+1}) (green and blue).

1.4.18. Substrate specificity of XAC0610 and other GGDEF proteins:

We already showed that XAC0610 is able to recognize 2´dGTP as a substrate. To see if this protein is able to recognize other possible substrates, we repeated the enzymatic assays using other naturally occurring nucleotide triphosphates such as: ATP, ITP, dATP and dITP. These experiments showed that among the substrates tested, only GTP and 2´dGTP acts as substrates for XAC0610 (Fig. 45).

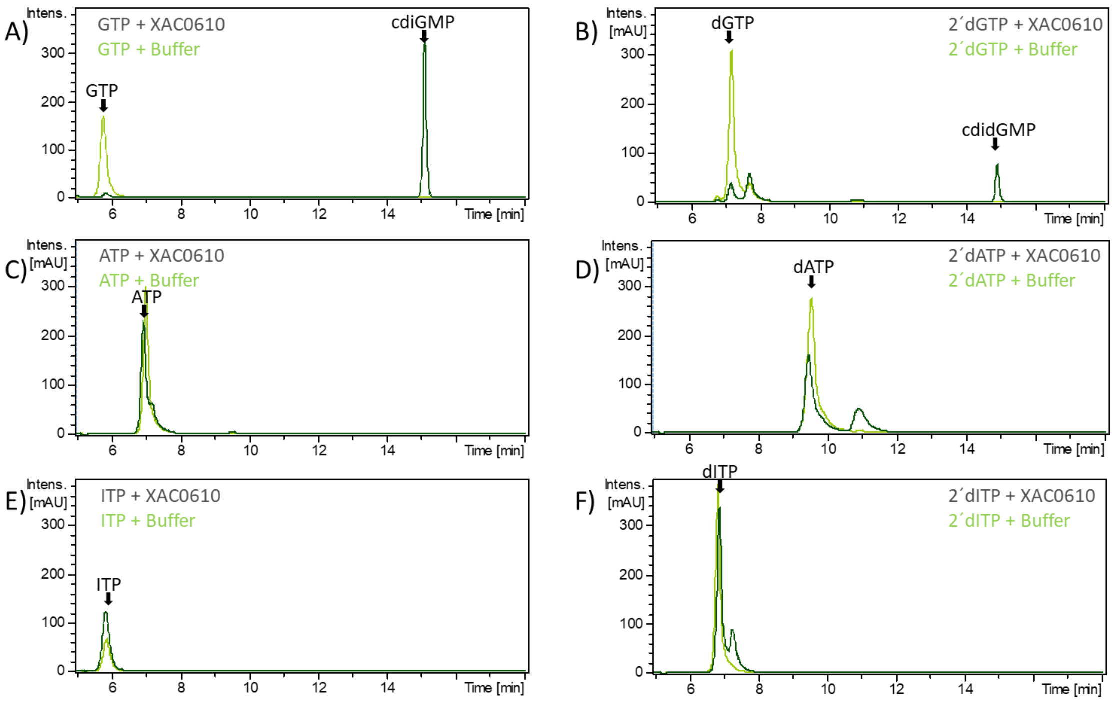


Figure 45: XAC0610 recognize GTP and 2´dGTP as substrates: HPLC-MS analysis of XAC0610 enzymatic reactions with different substrates. UV 259 nm chromatograms of enzymatic reactions that were done at 30 ºC with 5 µM XAC0610 in a volume reaction of 100 µl of reaction buffer (20 mM Tris-Cl pH 8.00, 5 mM MgCl2, 100 mM NaCl) and 1 mM of each tested substrate (A) GTP, B) 2´dGTP, C) ATP, D) 2`dATP, E) ITP or F) 2`dITP). The reactions were diluted after 1 hour by adding 900 ul of the reaction buffer and stored at -20 ºC until its analysis by HPLC-ESI-MS/MS. The products were analyzed by HPLC-ESI-MS/MS in a LC-10ADvp Shimadzu equipment using a C-18 column. The column was equilibrated with 2 % of buffer B (100 % methanol) and 98 % of buffer A (25 mM ammonium formiate, pH 7). The separation protocol was 2 % of buffer A up to 7 minutes, then the concentration of B was elevated up to 80 % within 17 minutes, this concentration was maintained up to 25 minutes, 80 % to 2 % up to 25 minutes and 2 % up to 50 minutes. The identity of the peaks was confirmed by inspection of the molecular mass in the ion chromatogram recorded by the Amazon Speed ETD (Bruker Daltonics) mass spectrometer equipped with an ESI interface.

To test if other diguanylate cyclase GGDEF domains could employ alternative NTPs as substrates, we purified (Fig. 46) and performed enzymatic assays using GGDEF domain containing proteins from different species (Fig. 47-52): *Leptospira interrogans* (LIC11128 and LIC11131), *Caulobacter Crescentus* (PleD), *Xanthomonas citri* (XAC0424 and XAC2810) and *Geobacter sulfurreducens* (GSU1658) employing GTP, 2´dGTP, ATP, 2´dATP, ITP, and 2´dITP as substrates. Purified proteins were incubated with 1 mmol/L of each substrate in separated tubes an incubated for 16 hours at 30 °C, HPLC technique was used to evaluate the formation of the products (see methods for details). All the tested proteins showed activity with GTP (Fig. 47-52). On the other hand, just GSU1658 showed activity with an NTP different to GTP (Fig. 47). GSU1658 is a well-characterized promiscuous enzyme previously showed to recognize GTP, ATP, and ITP (Hallberg et al. 2019); here, we show that beside these, GSU1658 can recognize 2´ dGTP but neither 2´dATP nor 2´dITP (Fig. 47).

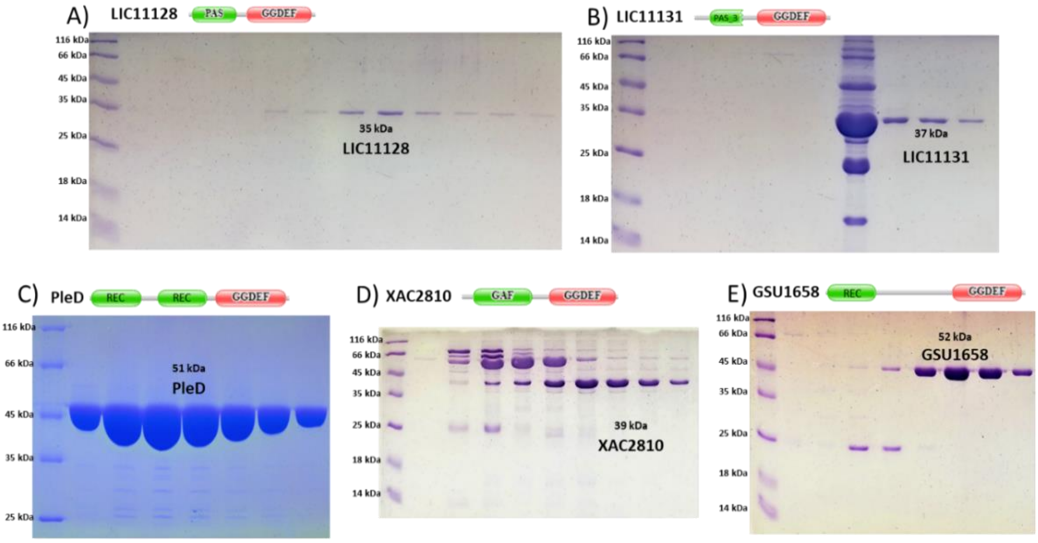


Figure 46: Purification of GGDEF domain containing proteins from different species: SDS-PAGE gels of the purification of 5 GGDEF proteins. Above each gel the scheme of the protein is showed. Two proteins belong to *Leptospira interrogans* (A and B), one to *Caulobacter crescentus* (C), one to *Xanthomonas citri* (D) and one to *Geobacter sulfurreducens* (E).

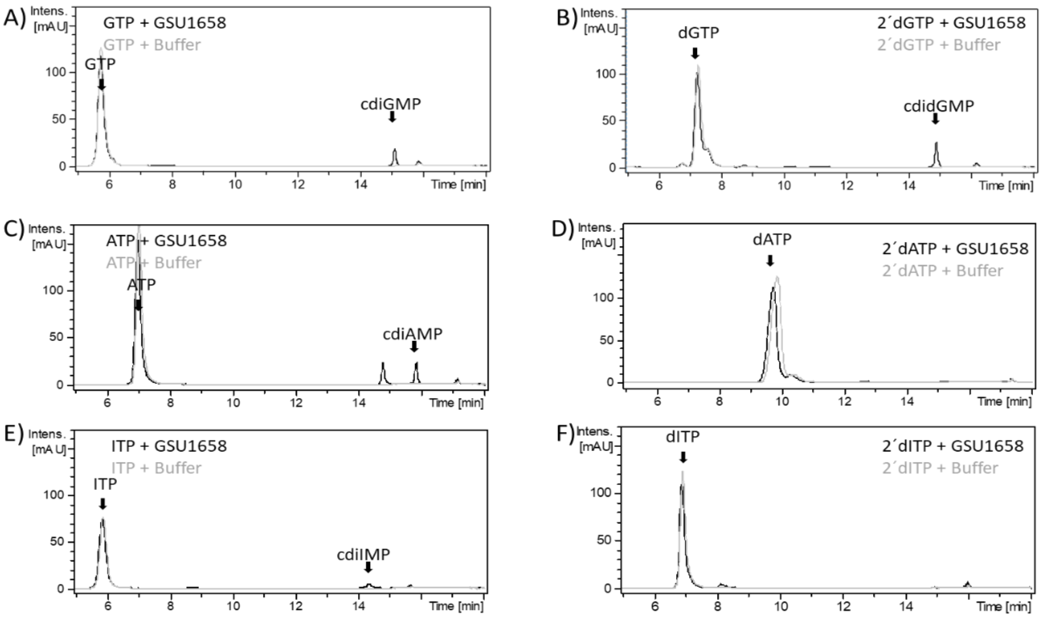


Figure 47: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 1): HPLC-MS analysis of GSU1658 enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

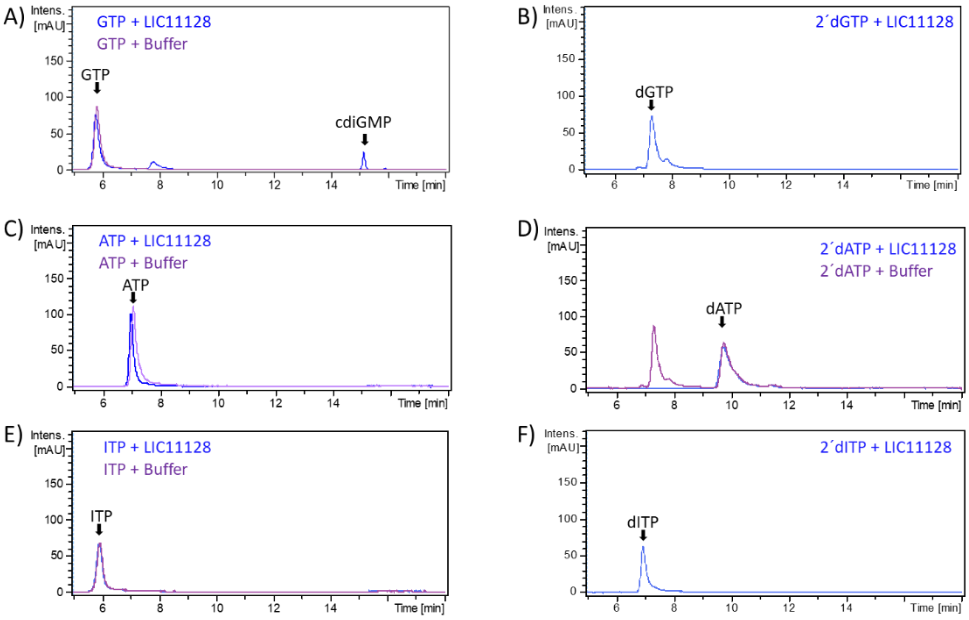


Figure 48: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 2): HPLC-MS analysis of LIC11128 enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

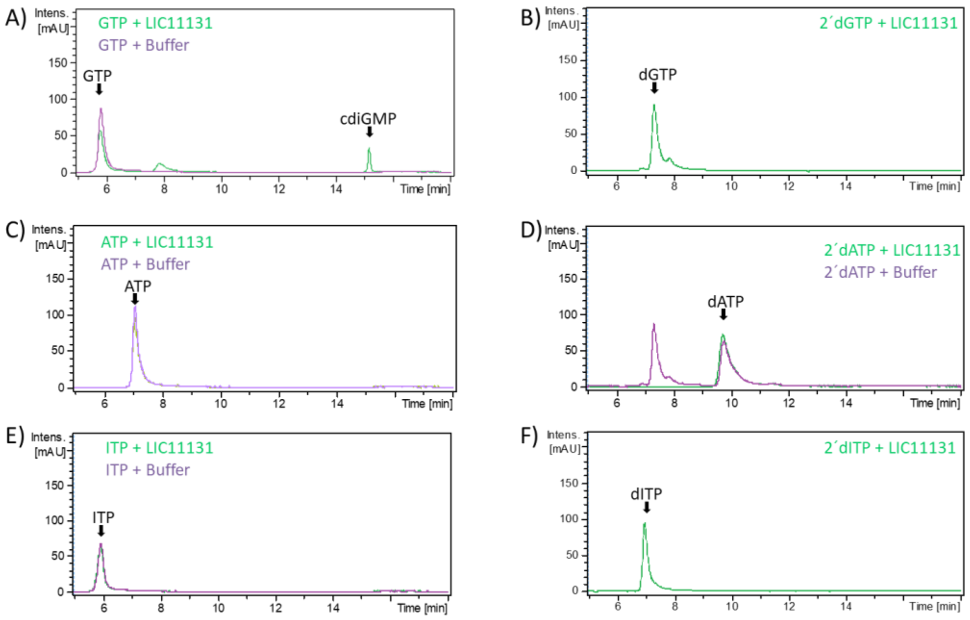


Figure 49: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 3): HPLC-MS analysis of LIC11131 enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

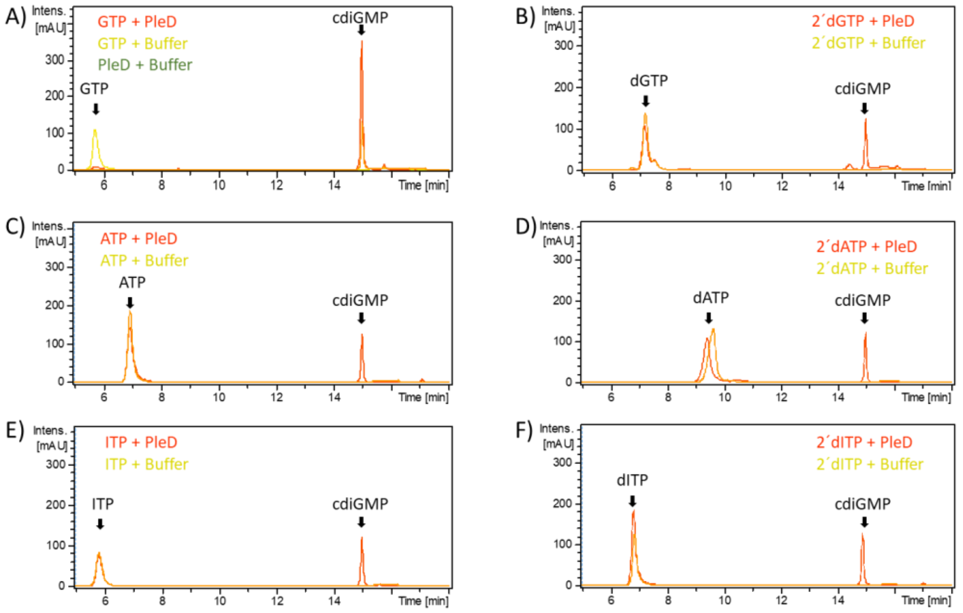


Figure 50: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 4): HPLC-MS analysis of PleD enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

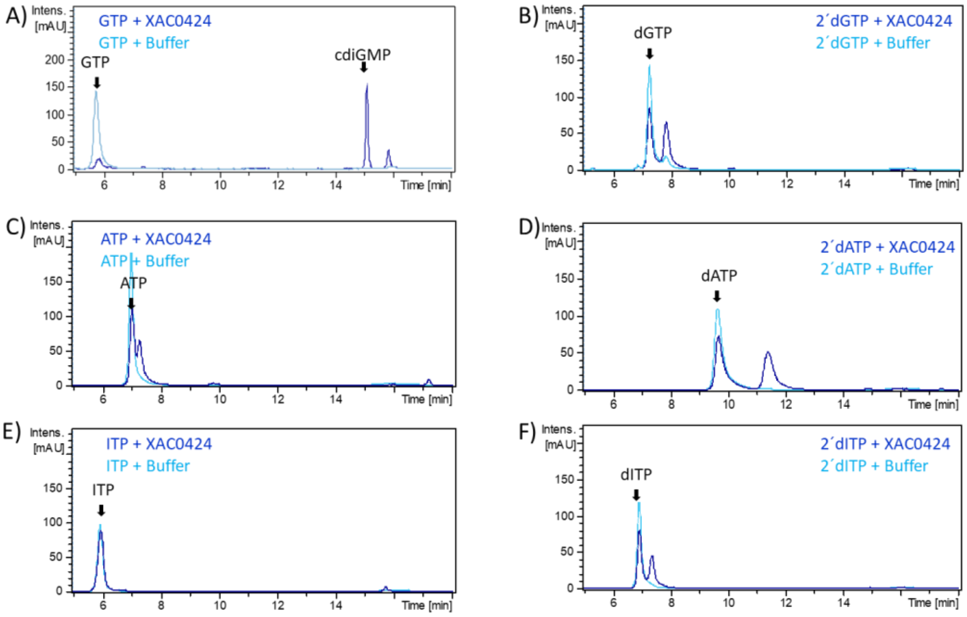


Figure 51: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 5): HPLC-MS analysis of XAC0424 enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

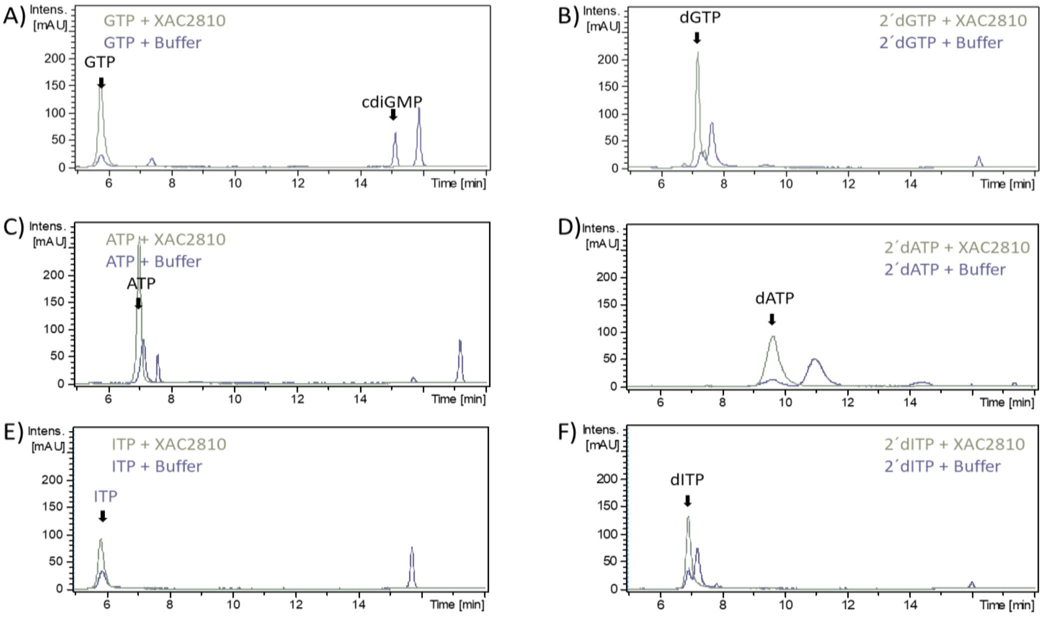


Figure 52: 2´dGTP is not a common substrate in GGDEF domain containing proteins (part 6): HPLC-MS analysis of XAC2810 enzymatic reactions with different substrates (GTP (A), 2´dGTP (B), ATP (C), 2`dATP (D), ITP (E) or 2`dITP (F)). UV 259 nm chromatograms of enzymatic reactions that were done as in Fig. 33.

1.4.19. GSU1658 enzymatic assays with combination of different substrates:

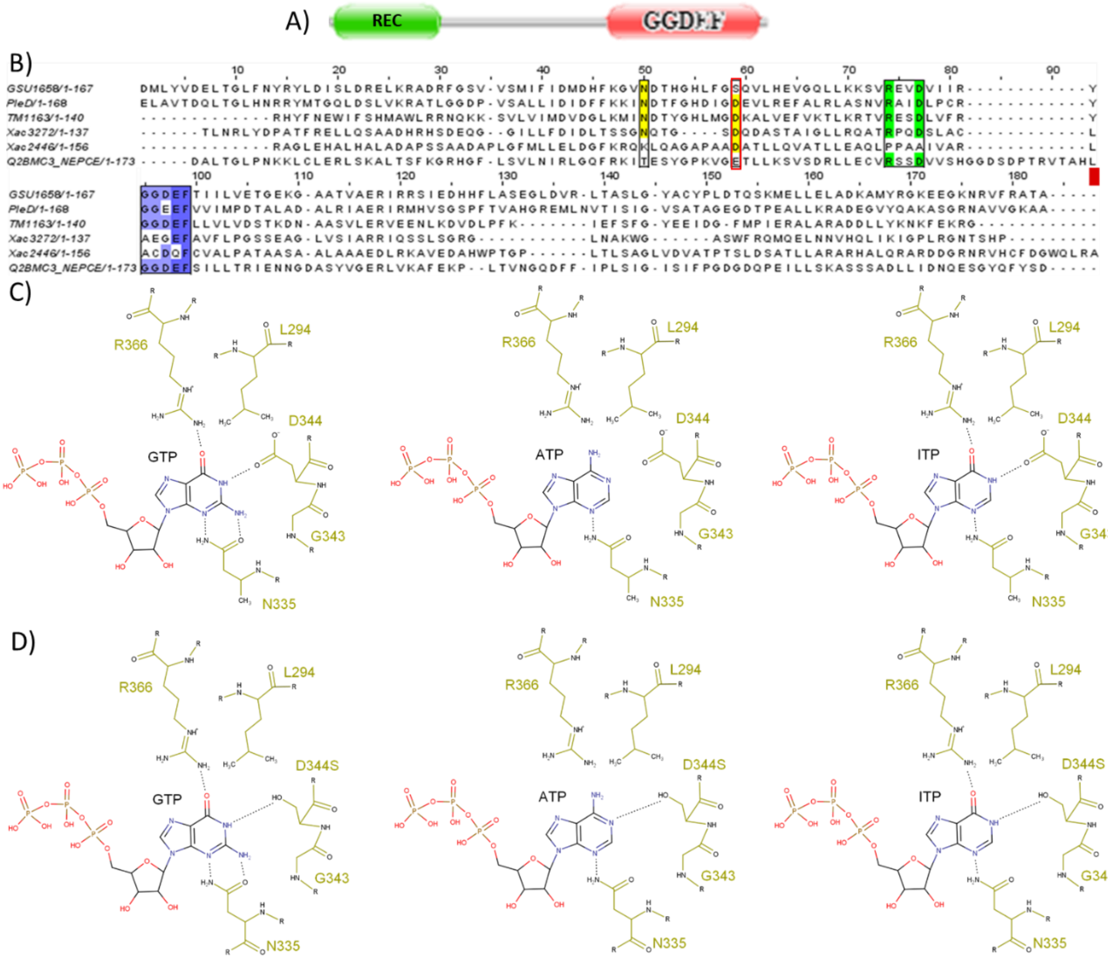

A)
REC
GGDEF
B)
10 20 30 40 50 60 70 80 90
GSU1658/1-167 DMLYVDELTGLFNYRYLDISLDRELKRADRFGSV-VSMIFIDMDHFKGVNDTHGHLFGSQVLHEVGQLLKKSVREVDVIIR------------Y
PleD/1-168 ELAVTDQLTGLHNRRYMTGQLDSLVKRATLGGDP-VSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRAIDLPCR------------Y
TM1163/1-140 -------------RHYFNEWIFSHMAWLRRNQKK-SVLVIMDVDGLKMINDTYGHLMGDKALVEFVKTLKRTVRESDLVFR------------Y
Xac3272/1-137 ------TLNRLYDPATFRELLQSAADHRHSDEQG--GILLFDIDLTSSGNQTG----SDQDASTAIGLLRQATRPQDSLAC------------L
Xac2446/1-156 -------------RAGLEHALHALADAPSSAADAPLGFMLLELDGFKRQKLQAGAPAADATLLQVATLLEAQLPPAAIVAR------------L
Q2BMC3_NEPCE/1-173 -----DALTGLPNKKLCLERLSKALTSFKGRHGF-LSVLNIRLGQFRKITESYGPKVGETLLKSVSDRLLECVRSSDVVSHGGDSDPTRVTAHL
100 110 120 130 140 150 160 170 180
GSU1658/1-167 GGDEFTIILVETGEKG-AATVAERIRRSIEDHHFLASEGLDVR-LTASLG-YACYPLDTQSKMELLELADKAMYRGKEEGKNRVFRATA-----
PleD/1-168 GGEEFVVIMPDTALAD-ALRIAERIRMHVSGSPFTVAHGREMLNVTISIG-VSATAGEGDTPEALLKRADEGVYQAKASGRNAVVGKAA-----
TM1163/1-140 GGDEFLLVLVDSTKDN-AASVLERVEENLKDVDLPFK-------IEFSFG-YEEIDG-FMPIERALARADDLLYKNKFEKRG------------
Xac3272/1-137 AEGEFAVFLPGSSEAG-LVSIARRIQSSLSGRG-----------LNAKWG------ASWFRQMQELNNVHQLIKIGPLRGNTSHP---------
Xac2446/1-156 ACDQFCVALPATAASA-ALAAAEDLRKAVEDAHWPTGP------LTLSAGLVDVATPTSLDSATLLARARHALQRARDDGRNRVHCFDGWQLRA
Q2BMC3_NEPCE/1-173 GGDEFSILLTRIENNGDASYVGERLVKAFEKP--LTVNGQDFF-IPLSIG-ISIFPGDGDQPEILLSKASSSADLLIDNQESGYQFYSD-----
C)
R366
L294
D344
GTP
G343
N335
R366
L294
D344
ATP
G343
N335
R366
L294
D344
ITP
G343
N335
D)
R366
L294
D344S
GTP
G343
N335
R366
L294
D344S
ATP
G343
N335
R366
L294
D344S
ITP
G343
N335

Figure 53. GSU1658 could recognize different substrates: A) GSU1658 domains architecture. B) Alignment of 6 GGDEF domains showing that GSU1658 present a serine substitution in the position of the conserved D344 (PleD numeration) residue. D) 2D chemical models of the nucleotide's recognition site showing that a canonical GGDEF domain has 3 residues (R366, D344 and N335 (PleD numeration)) forming hydrogen bonds with the guanine moiety of GTP. D344 is not able to form hydrogen bond with N1 of ATP, but a serine substitution in this position allows N1 of GTP and ATP to bind.

GSU1658 is a *Geobacter sulfurreducens* protein with two domains: An N-terminal REC domain and a C-terminal GGDEF domain (Fig. 53A). Hallberg et al., (2016) showed that GSU1658 is able to recognize ATP in its active site, probably due to a serine in the position corresponding to PleD D344 (Fig. 53B). A D344S substitution could lead to the formation of a hydrogen bond with N1 of ATP, while maintaining the possibility to bind GTP in the active site. Although this seems to be true, apparently that is not the only feature that determines the promiscuity of the GGDEF domain of GSU1658. For instance, Hallberg et al., (2016) performed enzymatic assays with PleD$_{D344S}$ mutant and showed that this substitution alone was not sufficient to change the specificity for the substrate.

On the other hand, inosine triphosphate (ITP) is another naturally occurring nucleotide triphosphate in all cells (Lin et al., 2001) generated by pyrophosphorylation or stepwise phosphorylation of IMP, an essential metabolite of purine biosynthesis and a precursor of both AMP and GMP. Models of the GGDEF recognition site (Fig. 53D) lead us to hypothesize that ITP could also be recognized by GGDEF domains. Similar to the recognition of guanine moiety, ITP could form at least three hydrogen bonds with R366, D344 and N335 residues in the recognition site of PleD. Due to the demonstrated promiscuity of GSU1658 and that it can forms cdiIMP from ITP we were interested in testing the possibility that other (not yet described) cyclic dinucleotides can be formed.

Therefore, we performed enzymatic assays with GSU1658 using ATP, GTP or ITP as a substrate as well combinations of these substrates and product formation was evaluated by HPLC-MS (see Methodologies). As already shown, GSU1658 is able to form cdiGMP, cdiAMP, cdiIMP, and cGAMP (Fig. 47). Peak products showed the expected masses (690.1 g/ mol for cdiGMP, 658.1 g/mol for cdiAMP and 674.1 g/mol for cGAMP) of these three cyclic dinucleotides. Interestingly, when we mix ATP and ITP or GTP and ITP as substrates we obtain peaks with expected masses for cGIMP (675.1 g/mol) and cAIMP (659.1 g/mol) (Fig. 54).

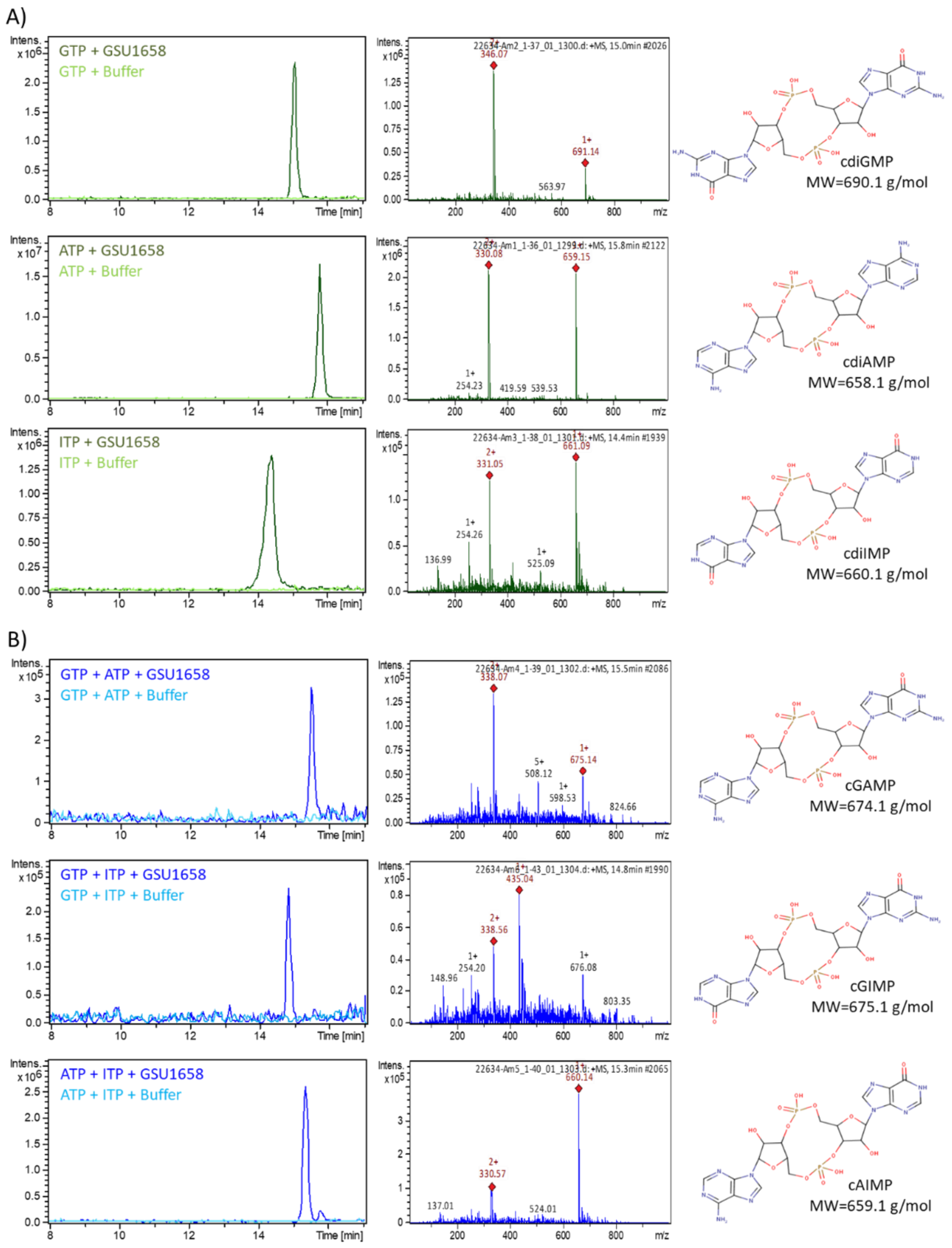


Figure 54. GSU1658 enzymatic assays with different substrates: HPLC-MS analysis of GSU1658 enzymatic reactions with different substrates, panels in the left show extracted ion chromatograms corresponding to the mass of the expected products. In the middle we showed the mass spectra of the highest peak in the chromatogram and in the right, we showed the molecular configuration of the expected products and its expected molecular mass. Enzymatic reactions were done at 30 ºC with 10 uM GSU1658 in a volume reaction of 100 ul containing 20 mM Tris-Cl pH 8.00, 5 mM MgCl2, 100 mM NaCl and 1 mM of each substrate. The reactions were quenched at 1 hour adding 900 ul of stop solution (10 mM HCl). HPLC was done as in Fig. 33. E) Reactions of GSU1658 using only one substrate (ATP, GTP or ITP). F) Reactions of GSU1658 mixing two different substrates.

To our knowledge, this is the first description of the enzymatic production of these new cyclic dinucleotides. dITP is potentially mutagenic and the level of this nucleotide is controlled by inosine triphosphate pyrophosphatases (Burgis & Cunningham, 2007; Savchenko et al., 2007). It is interesting to speculate that cdiIMP, cGIMP or cAIMP could act as a signaling molecule when ITP/dITP concentration in the cells increases beyond normal levels.

1.4.20. Crystallization tests of GSU1658:

In the case of GSU1658, we showed that this protein is able to recognize several substrates including GTP, ATP, ITP and 2´dGTP. To understand the structural basis of these recognitions we aimed to solve the structure of this protein in complex with those substrates. Recently, a crystal structure of the GGDEF domain of this protein was solved in complex with GTP (Hallberg et al., 2019).

First, we tested 6 different crystallization kits with a full-length version of GSU1658 with a 6xHis C-terminal tag. We did not obtain crystals in these assays. Since several crystal contacts in PleD crystals are formed by its REC domains and that the GGDEF domain of GSU1658 has around 38% identity with the GGDEF domain of PleD (Fig. 55C), we reasoned that a chimera where the GSU1658 GGDEF domain has been attached to the REC domains of PleD (Fig. 55A and 55B) could help to crystallize the GSU1658 GGDEF domain. We therefore cloned, expressed, and purified a protein that contains residues 1-293 of PleD followed by residues 297-458 of GSU1658 (Fig. 55D). Unfortunately, attempts to crystallize these constructions using similar crystallization conditions of PleD (glycine 1 M, PEG 20000 13,0 – 16,0 %, dioxane 1,0 – 2,5 %, pH 8.2 – 9.2) were not successful.

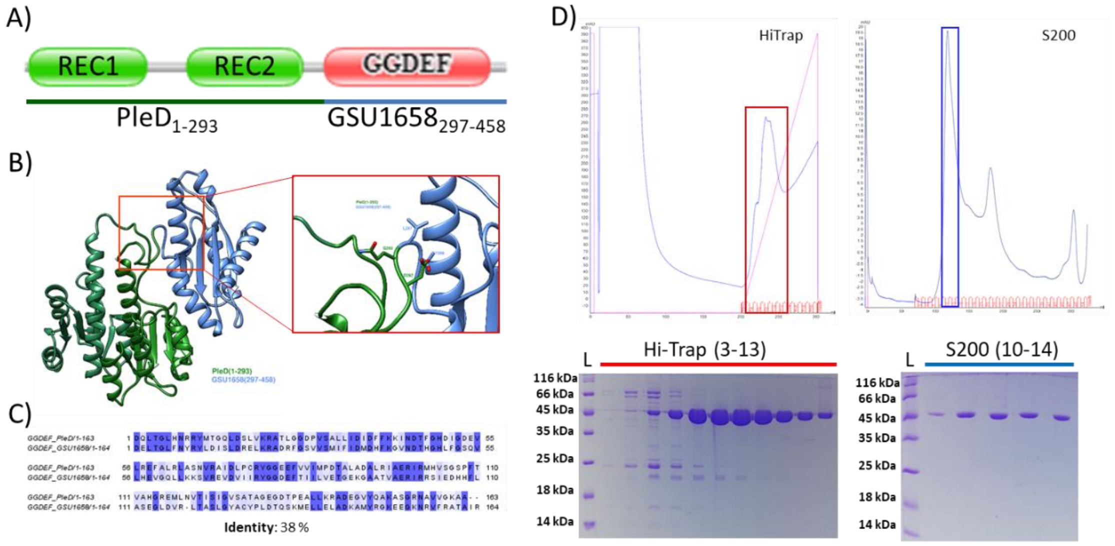


Figure 55: PleD$_{1-293}$-GSU1658$_{297-458}$ chimera purification: A) Scheme of the construction that was purified, this construction has the aminoacids 1 to 293 of PleD covering the two REC domains and the aminoacids 297 to 458 of GSU1658 corresponding to the GGDEF domain of this protein. B) Model produced by Swiss model (https://swissmodel.expasy.org/) of the produced chimera. In green we show the region corresponding to aminoacids from PleD and in blue aminoacids from GSU1658. The panel in the right shows a zoom of the connection loop between the two proteins. C) An alignment of the GGGDEF domains of PleD and GSU1658 colored according to the degree of conservation. D) In the left we show a nickel affinity chromatography result with an SDS-PAGE (below) of the corresponding fractions and in the right, we show the result of a size exclusion chromatography of the concentrated fractions of the affinity chromatography.

1.5.- CONCLUSIONS:

Seven single mutations on the active site of the GGDEF-containing diguanylate cyclase PleD were tested in crystallization assays in an attempt to study the effects of the mutations in the structure of the active site. The structure of two mutants (N335T and D344N) was solved in an inactive conformational state with the GGDEF domains crosslinked with the Rec1 domain by two c-di-GMP molecules. The overall structure of the active sites of these two mutants was very similar to that of the wild-type protein but no electron density corresponding to c-di-GMP was observed. This indicates that these two changes abolished the binding of the guanine moiety in the active site.

Another GGDEF containing-protein studied here was the multidomain protein XAC0610. Several different constructions were created based on different combinations of domains and several of them were purified. However, none of the tested constructions produced crystals and therefore no structure of this protein is available. On the other hand, five single mutants in the active site of the GGDEF domain of this protein were constructed and the capacity to use ATP or GTP as substrates was tested. Just one mutation (R793H, corresponding to R366H in PleD) conserved the wild-type activity but none of the mutants gained the capacity to recognize ATP. It is intriguingly why R793 is a very conserved residue, but the *in vitro* activity is not affected by its substitution to histidine.

PleD, XAC0610 and another five GGDEF-containing proteins were tested for their capacity to recognize other natural occurring nucleotide triphosphates as a substrate. A GGDEF-containing protein GSU1658 from *Geobacter sulfurreduscens* was shown to produce the cyclic dinucleotides and hybrid cyclic dinucleotides: c-di-GMP, c-di-AMP, c-di-IMP, c-di-2´dGMP, c-GAMP, c-GIMP, and c-AIMP. Interestingly, XAC0610 recognized both 2´dGTP and GTP as substrates. Enzymatic kinetics experiments in XAC0610 using the two substrates showed that the the cyclic product cdiGMP and the linear product pppGp2´dG are formed in similar amounts when the GTP and 2´dGTP are present at similar concentrations, opening up the possibility that, under certain conditions, pppGpdG could accumulate. Whether pppGpdG has a physiological role in bacterial physiology or is a dead-end side-reactions remains to be determined. The enzyme kinetic modeling also showed that the observed ratios of cdiGMP and pppGpdG formed is due, in part, to the different substrate preference depending on the state of the monomer that composed the XAC0610 dimer, specifically, the capacity of the empty XAC0610 monomer to discriminate if the other monomer has GTP or dGTP. Conversion of the linear product to the cyclic cGMP-dGMP product is extremely slow due to reduced nucleophilicity of 3´-OH of 2´dGTP on the α-phosphate of GTP. Similar questions are raised by the observation of the production of cGIMP and cAIMP.

1.6.- REFERENCES:

Ablasser, A., Goldeck, M., Cavlar, T., Deimling, T., Witte, G., Röhl, I., Hopfner, K. P., Ludwig, J., & Hornung, V. (2013). CGAS produces a 2′-5′-linked cyclic dinucleotide second messenger that activates STING. *Nature*, *498*(7454), 380–384. https://doi.org/10.1038/nature12306

Aldridge, P., Paul, R., Goymer, P., Rainey, P., & Jenal, U. (2003). Role of the GGDEF regulator PleD in polar development of Caulobacter crescentus. *Molecular Microbiology*, *47*(6), 1695–1708. https://doi.org/10.1046/j.1365-2958.2003.03401.x

Almblad, H., Randall, T. E., Liu, F., Leblanc, K., Groves, R. A., Kittichotirat, W., Winsor, G. L., Fournier, N., Au, E., Groizeleau, J., Rich, J. D., Lou, Y., Granton, E., Jennings, L. K., Singletary, L. A., Winstone, T. M. L., Good, N. M., Bumgarner, R. E., Hynes, M. F., … Harrison, J. J. (2021). Bacterial cyclic diguanylate signaling networks sense temperature. *Nature Communications*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-22176-2

Amikam, D., & Galperin, M. Y. (2006). PilZ domain is part of the bacterial c-di-GMP binding protein. *Bioinformatics*, *22*(1), 3–6. https://doi.org/10.1093/bioinformatics/bti739

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, *2*. https://doi.org/10.1038/msb4100050

Bai, Y., Yang, J., Eisele, L. E., Underwood, A. J., Koestler, B. J., Waters, C. M., Metzger, D. W., & Bai, G. (2013). Two DHH subfamily 1 proteins in Streptococcus pneumoniae possess cyclic Di-AMP phosphodiesterase activity and affect bacterial growth and virulence. *Journal of Bacteriology*, *195*(22), 5123–5132. https://doi.org/10.1128/JB.00769-13

Bai, Y., Yang, J., Zhou, X., Ding, X., Eisele, L. E., & Bai, G. (2012). Mycobacterium tuberculosis Rv3586 (DacA) is a diadenylate cyclase that converts ATP or ADP into c-di-amp. *PLoS ONE*, *7*(4), 1–10. https://doi.org/10.1371/journal.pone.0035206

Barnes, C. O., Wu, Y., Song, J., Lin, G., Baxter, E. L., Brewster, A. S., Nagarajan, V., Holmes, A., Michael Soltis, S., Sauter, N. K., Ahn, J., Cohen, A. E., & Calero, G. (2019). The crystal structure of dGTPase reveals the molecular basis of dGTP selectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(19), 9333–9339. https://doi.org/10.1073/pnas.1814999116

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R., & Leslie, A. G. W. (2011). iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 271–281. https://doi.org/10.1107/S0907444910048675

Bobrov, A. G., Kirillina, O., & Perry, R. D. (2005). The phosphodiesterase activity of the HmsP EAL domain is required for negative regulation of biofilm formation in Yersinia pestis. *FEMS Microbiology Letters*, *247*(2), 123–130. https://doi.org/10.1016/j.femsle.2005.04.036

Buckstein, M. H., He, J., & Rubin, H. (2008). Characterization of nucleotide pools as a function of physiological state in Escherichia coli. *Journal of Bacteriology*, *190*(2), 718–726. https://doi.org/10.1128/JB.01020-07

Burgis, N. E., & Cunningham, R. P. (2007). Substrate specificity of RdgB protein, a deoxyribonucleoside triphosphate pyrophosphohydrolase. *Journal of Biological Chemistry*, *282*(6), 3531–3538. https://doi.org/10.1074/jbc.M608708200

Chan, C., Paul, R., Samoray, D., Amiot, N. C., Giese, B., Jenal, U., & Schirmer, T. (2004). Structural basis of activity and allosteric control of diguanylate cyclase. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(49), 17084–17089. https://doi.org/10.1073/pnas.0406134101

Christen, M., Christen, B., Folcher, M., Schauerte, A., & Jenal, U. (2005). Identification and characterization of a cyclic di-GMP-specific phosphodiesterase and its allosteric control by GTP. *Journal of Biological Chemistry*, *280*(35), 30829–30837. https://doi.org/10.1074/jbc.M504429200

Cohen, D., Melamed, S., Millman, A., Shulman, G., Oppenheimer-Shaanan, Y., Kacen, A., Doron, S., Amitai, G., & Sorek, R. (2019). Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature*, *574*(7780), 691–695. https://doi.org/10.1038/s41586-019-1605-5

Corrigan, R. M., Abbott, J. C., Burhenne, H., Kaever, V., & Gründling, A. (2011). C-di-amp is a new second messenger in staphylococcus aureus with a role in controlling cell size and envelope stress. *PLoS Pathogens*, *7*(9). https://doi.org/10.1371/journal.ppat.1002217

Corrigan, R. M., Campeotto, I., Jeganathan, T., Roelofs, K. G., Lee, V. T., & Gründling, A. (2013). Systematic identification of conserved bacterial c-di-AMP receptor proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 9084–9089. https://doi.org/10.1073/pnas.1300595110

Corrigan, R. M., & Gründling, A. (2013). Cyclic di-AMP: Another second messenger enters the fray. *Nature Reviews Microbiology*, *11*(8), 513–524. https://doi.org/10.1038/nrmicro3069

Crossman, L. C., Spiro, S., He, Y., Zhang, L., Heeb, S., Cámara, M., Williams, P., Maxwell, J., Ryan, R. P., Fouhy, Y., Lucey, J. F., Crossman, L. C., Spiro, S., He, Y., Zhang, L., Williams, P., Dow, J. M., Heeb, S., & Ca, M. (2017). Erratum: Cell-cell signaling in Xanthomonas campestris involves an HD-GYP domain protein that functions in cyclic diGMP turnover (Proceedings of the National Academy of Sciences of the United States of America (2016) 103 (6712-6717) DOI: 10.1073/pnas.0600. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(33), E7031. https://doi.org/10.1073/pnas.1712524114

Davies, B. W., Bogard, R. W., Young, T. S., & Mekalanos, J. J. (2012). Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for V. cholerae virulence. *Cell*, *149*(2), 358–370. https://doi.org/10.1016/j.cell.2012.01.053

Diner, E. J., Burdette, D. L., Wilson, S. C., Monroe, K. M., Kellenberger, C. A., Hyodo, M., Hayakawa, Y., Hammond, M. C., & Vance, R. E. (2013). The Innate Immune DNA Sensor cGAS Produces a Noncanonical Cyclic Dinucleotide that Activates Human STING. *Cell Reports*, *3*(5), 1355–1361. https://doi.org/10.1016/j.celrep.2013.05.009

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, *66*(4), 486–501. https://doi.org/10.1107/S0907444910007493

Ferraro, P., Franzolin, E., Pontarin, G., Reichard, P., & Bianchi, V. (2009). Quantitation of cellular deoxynucleoside triphosphates. *Nucleic Acids Research*, *38*(6). https://doi.org/10.1093/nar/gkp1141

Hallberg, Z. F., Chan, C. H., Wright, T. A., Kranzusch, P. J., Doxzen, K. W., Park, J. J., Bond, D. R., & Hammond, M. C. (2019). Structure and mechanism of a hypr GGDEF enzyme that activates cGAMP signaling to control extracellular metal respiration. *ELife*, *8*, 1–36. https://doi.org/10.7554/eLife.43959

Hallberg, Z. F., Wang, X. C., Wright, T. A., Nan, B., Ad, O., Yeo, J., & Hammond, M. C. (2016). Hybrid promiscuous (Hypr) GGDEF enzymes produce cyclic AMP-GMP (3', 3'-cGAMP). *Proceedings of the National Academy of Sciences of the United States of America*, *113*(7), 1790–1795. https://doi.org/10.1073/pnas.1515287113

He, J., Ruan, W., Sun, J., Wang, F., & Yan, W. (2018). Functional characterization of c-di-GMP signaling-related genes in the probiotic Lactobacillus acidophilus. *Frontiers in Microbiology*, *9*(AUG), 1–15. https://doi.org/10.3389/fmicb.2018.01935

Heikaus, C. C., Pandit, J., & Klevit, R. E. (2009). Cyclic Nucleotide Binding GAF Domains from Phosphodiesterases: Structural and Mechanistic Insights. *Structure*, *17*(12), 1551–1557. https://doi.org/10.1016/j.str.2009.07.019

Hickman, J. W., & Harwood, C. S. (2008). Identification of FleQ from Pseudomonas aeruginosa as a c-di-GMP-responsive transcription factor. *Molecular Microbiology*, *69*(2), 376–389. https://doi.org/10.1111/j.1365-2958.2008.06281.x

Ho, Y. S. J., Burden, L. M., & Hurley, J. H. (2000). Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *EMBO Journal*, *19*(20), 5288–5299. https://doi.org/10.1093/emboj/19.20.5288

Huber, H. E., Beauchamp, B. B., & Richardson, C. C. (1988). Escherichia coli dGTP triphosphohydrolase is inhibited by gene 1.2 protein of bacteriophage T7. *Journal of Biological Chemistry*, *263*(27), 13549–13556. https://doi.org/10.1016/s0021-

9258(18)68277-8

Jenal, U., & Malone, J. (2006). Mechanisms of cyclic-di-GMP signaling in bacteria. *Annual Review of Genetics*, *40*, 385–407. https://doi.org/10.1146/annurev.genet.40.110405.090423

Kellenberger, C. A., Chen, C., Whiteley, A. T., Portnoy, D. A., & Hammond, M. C. (2015). RNA-Based Fluorescent Biosensors for Live Cell Imaging of Second Messenger Cyclic di-AMP. *Journal of the American Chemical Society*, *137*(20), 6432–6435. https://doi.org/10.1021/jacs.5b00275

Kellenberger, C. A., Wilson, S. C., Hickey, S. F., Gonzalez, T. L., Su, Y., Hallberg, Z. F., Brewer, T. F., Iavarone, A. T., Carlson, H. K., Hsieh, Y. F., & Hammond, M. C. (2015). GEMM-I riboswitches from Geobacter sense the bacterial second messenger cyclic AMP-GMP. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5383–5388. https://doi.org/10.1073/pnas.1419328112

Kellenberger, C. A., Wilson, S. C., Sales-Lee, J., & Hammond, M. C. (2013). RNA-Based Fluorescent Biosensors for Live Cell Imaging of Second Messengers Cyclic di-GMP and Cyclic AMP-GMP. *Journal of the American Chemical Society*, *135*(13), 4906–4909. https://doi.org/10.1021/ja311960g

Kranzusch, P. J., Lee, A. S. Y., Wilson, S. C., Solovykh, M. S., Vance, R. E., Berger, J. M., & Doudna, J. A. (2014). Structure-guided reprogramming of human cgas dinucleotide linkage specificity. *Cell*, *158*(5), 1011–1021. https://doi.org/10.1016/j.cell.2014.07.028

Leduc, J. L., & Roberts, G. P. (2009). Cyclic di-GMP allosterically inhibits the CRP-like protein (Clp) of Xanthomonas axonopodis pv. citri. *Journal of Bacteriology*, *191*(22), 7121–7122. https://doi.org/10.1128/JB.00845-09

Lin, S., McLennan, A. G., Ying, K., Wang, Z., Gu, S., Jin, H., Wu, C., Liu, W., Yuan, Y., Tang, R., Xie, Y., & Mao, Y. (2001). Cloning, Expression, and Characterization of a Human Inosine Triphosphate Pyrophosphatase Encoded by the ITPA Gene. *Journal of Biological Chemistry*, *276*(22), 18695–18701. https://doi.org/10.1074/jbc.M011084200

Martinez, S. E., Bruder, S., Schultz, A., Zheng, N., Schultz, J. E., Beavo, J. A., & Linder, J. U. (2005). Crystal structure of the tandem GAF domains from a cyanobacterial adenylyl cyclase: Modes of ligand binding and dimerization. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(8), 3082–3087. https://doi.org/10.1073/pnas.0409913102

Martinez, S. E., Wu, A. Y., Glavas, N. A., Tang, X. B., Turley, S., Hol, W. G. J., & Beavo, J. A. (2002). The two GAF domains in phosphodiesterase 2A have distinct roles in dimerization and in cGMP binding. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13260–13265. https://doi.org/10.1073/pnas.192374899

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography*, *40*(4), 658–674. https://doi.org/10.1107/S0021889807021206

Muradov, K. G., Boyd, K. K., Martinez, S. E., Beavo, J. A., & Artemyev, N. O. (2003). The GAFa domains of rod cGMP-phosphodiesterase 6 determine the selectivity of the enzyme dimerization. *Journal of Biological Chemistry*, *278*(12), 10594–10601. https://doi.org/10.1074/jbc.M208456200

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 355–367. https://doi.org/10.1107/S0907444911001314

Nakai, H., & Richardson, C. C. (1990). The gene 1.2 protein of bacteriophage T7 interacts with the Escherichia coli dGTP triphosphohydrolase to form a GTP-binding protein. *Journal of Biological Chemistry*, *265*(8), 4411–4419. https://doi.org/10.1016/s0021-9258(19)39580-8

Nelson, J. W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J. X., & Breaker, R. R. (2013). Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nature Chemical Biology*, *9*(12), 834–839. https://doi.org/10.1038/nchembio.1363

Oliveira, M. C., Teixeira, R. D., Andrade, M. O., Pinheiro, G. M. S., Ramos, C. H. I., & Farah, C. S. (2015). Cooperative substrate binding by a diguanylate cyclase. *Journal of Molecular Biology*, *427*(2), 415–432. https://doi.org/10.1016/j.jmb.2014.11.012

Paul, R., Abel, S., Wassmann, P., Beck, A., Heerklotz, H., & Jenal, U. (2007). Activation of the diguanylate cyclase PleD by phosphorylation-mediated dimerization. *Journal of Biological Chemistry*, *282*(40), 29170–29177.

https://doi.org/10.1074/jbc.M704702200

Paul, R., Weiser, S., Amiot, N. C., Chan, C., Schirmer, T., Giese, B., & Jenal, U. (2004). Cell cycle-dependent dynamic localization of a bacterial response regulator with a novel di-guanylate cyclase output domain. *Genes and Development*, *18*(6), 715–727. https://doi.org/10.1101/gad.289504

Rao, F., See, R. Y., Zhang, D., Toh, D. C., Ji, Q., & Liang, Z. X. (2010). YybT is a signaling protein that contains a cyclic dinucleotide phosphodiesterase domain and a GGDEF domain with ATPase activity. *Journal of Biological Chemistry*, *285*(1), 473–482. https://doi.org/10.1074/jbc.M109.040238

Ross, P., Weinhouse, H., Aloni, Y., Michaeli, D., Weinberger-Ohana, P., Mayer, R., Braun, S., De Vroom, E., Van Der Marel, G. A., Van Boom, J. H., & Benziman, M. (1987). Regulation of cellulose synthesis in Acetobacter xylinum by cyclic diguanylic acid. In *Nature* (Vol. 325, Issue 6101, pp. 279–281). https://doi.org/10.1038/325279a0

Roy, A., Petrova, O., & Sauer, K. (2013). Extraction and Quantification of Cyclic Di-GMP from Pseudomonas aeruginosa. *Bio-Protocol*, *3*(14). https://doi.org/10.21769/bioprotoc.828

Ryjenkov, D. A., Tarutina, M., Moskvin, O. V., & Gomelsky, M. (2005). Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: Insights into biochemistry of the GGDEF protein domain. *Journal of Bacteriology*, *187*(5), 1792–1798. https://doi.org/10.1128/JB.187.5.1792-1798.2005

Savchenko, A., Proudfoot, M., Skarina, T., Singer, A., Litvinova, O., Sanishvili, R., Brown, G., Chirgadze, N., & Yakunin, A. F. (2007). Molecular Basis of the Antimutagenic Activity of the House-Cleaning Inosine Triphosphate Pyrophosphatase RdgB from Escherichia coli. *Journal of Molecular Biology*, *374*(4), 1091–1103. https://doi.org/10.1016/j.jmb.2007.10.012

Schirmer, T. (2016). C-di-GMP Synthesis: Structural Aspects of Evolution, Catalysis and Regulation. *Journal of Molecular Biology*, *428*(19), 3683–3701. https://doi.org/10.1016/j.jmb.2016.07.023

Seto, D., Bhatnagar, S. K., & Bessman, M. J. (1988). The purification and properties of deoxyguanosine triphosphate triphosphohydrolase from Escherichia coli. *Journal of Biological Chemistry*, *263*(3), 1494–1499. https://doi.org/10.1016/s0021-9258(19)57330-6

Sun, L., Wu, J., Du, F., Chen, X., & Chen, Z. J. (2013). Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science*, *339*(6121), 786–791. https://doi.org/10.1126/science.1232458

Sun, Y., Liu, Y., Liu, X., Dang, X., Dong, X., & Xie, Z. (2020). Azorhizobium caulinodans c-di-GMP phosphodiesterase Chp1 involved in motility, EPS production, and nodulation of the host plant. *Applied Microbiology and Biotechnology*, *104*(6), 2715–2729. https://doi.org/10.1007/s00253-020-10404-6

Sureka, K., Choi, P. H., Precit, M., Delince, M., Pensinger, D. A., Huynh, T. A. N., Jurado, A. R., Goo, Y. A., Sadilek, M., Iavarone, A. T., Sauer, J. D., Tong, L., & Woodward, J. J. (2014). The cyclic dinucleotide c-di-AMP is an allosteric regulator of metabolic enzyme function. *Cell*, *158*(6), 1389–1401. https://doi.org/10.1016/j.cell.2014.07.046

Teixeira, R. D., Guzzo, C. R., Arévalo, S. J., Andrade, M. O., Abrahão, J., De Souza, R. F., & Farah, C. S. (2018). A bipartite periplasmic receptor– diguanylate cyclase pair (xac2383–xac2382) in the bacterium xanthomonas citri. *Journal of Biological Chemistry*, *293*(27), 10767–10781. https://doi.org/10.1074/jbc.RA118.003475

Traut, T. W. (1994). Physiological concentrations of purines and pyrimidines. *Molecular and Cellular Biochemistry*, *140*(1), 1–22. https://doi.org/10.1007/BF00928361

Tucker, C. L., Hurley, J. H., Miller, T. R., & Hurley, J. B. (1998). Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(11), 5993–5997. https://doi.org/10.1073/pnas.95.11.5993

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wang, H., Robinson, H., & Ke, H. (2010). Conformation changes, N-terminal involvement, and cGMP signal relay in the phosphodiesterase-5 GAF domain. *Journal of Biological Chemistry*, *285*(49), 38149–38156. https://doi.org/10.1074/jbc.M110.141614

Wassmann, P., Chan, C., Paul, R., Beck, A., Heerklotz, H., Jenal, U., & Schirmer, T. (2007). Structure of BeF3--Modified Response Regulator PleD: Implications for Diguanylate

Cyclase Activation, Catalysis, and Feedback Inhibition. *Structure*, *15*(8), 915–927. https://doi.org/10.1016/j.str.2007.06.016

Whitney, J. C., Colvin, K. M., Marmont, L. S., Robinson, H., Parsek, M. R., & Howell, P. L. (2012). Structure of the cytoplasmic region of PelD, a degenerate diguanylate cyclase receptor that regulates exopolysaccharide production in Pseudomonas aeruginosa. *Journal of Biological Chemistry*, *287*(28), 23582–23593. https://doi.org/10.1074/jbc.M112.375378

Witte, G., Hartung, S., Büttner, K., & Hopfner, K. P. (2008). Structural Biochemistry of a Bacterial Checkpoint Protein Reveals Diadenylate Cyclase Activity Regulated by DNA Recombination Intermediates. *Molecular Cell*, *30*(2), 167–178. https://doi.org/10.1016/j.molcel.2008.02.020

Yang, J., Bai, Y., Zhang, Y., Gabrielle, V. D., Jin, L., & Bai, G. (2014). Deletion of the cyclic di-AMP phosphodiesterase gene (cnpB) in Mycobacterium tuberculosis leads to reduced virulence in a mouse model of infection. *Molecular Microbiology*, *93*(1), 65–79. https://doi.org/10.1111/mmi.12641

Zähringer, F., Lacanna, E., Jenal, U., Schirmer, T., & Boehm, A. (2013). Structure and signaling mechanism of a zinc-sensory diguanylate cyclase. *Structure*, *21*(7), 1149–1157. https://doi.org/10.1016/j.str.2013.04.026

Zhu, D., Wang, L., Shang, G., Liu, X., Zhu, J., Lu, D., Wang, L., Kan, B., Zhang, J. ren, & Xiang, Y. (2014). Structural Biochemistry of a Vibrio cholerae Dinucleotide Cyclase Reveals Cyclase Activity Regulation by Folates. *Molecular Cell*, *55*(6), 931–937.

https://doi.org/10.1016/j.molcel.2014.08.001

Zoraghi, R., Corbin, J. D., & Francis, S. H. (2004). Properties and Functions of GAF Domains in Cyclic Nucleotide Phosphodiesterases and Other Proteins. Molecular Pharmacology, 65(2), 267–278. https://doi.org/10.1124/mol.65.2.267

CHAPTER 2. STUDIES IN THE CYANIDE METABOLISM OF BACILLUS SPP.

2.1.- INTRODUCTION

2.1.1. Chemistry of Cyanide

Hydrogen cyanide was once abundant in the earth´s atmosphere and is thought to have been an important reagent in the formation of biological molecules such as amino acids and nucleosides in the prebiotic earth (Ferus et al., 2020; Menor Salván et al., 2020; Todd & Öberg, 2020). Hydrogen cyanide is volatile while the cyanide ion is stable in aqueous solution. The pKa of the system $CN^- + H^+$ <-> HCN is 9.2 at 25 ^oC but it can vary depending on temperature and ionic strength: lower temperatures and lower ionic strength increase the pKa (Johnson, 2015).

Cyanide forms very stable complexes with many metals (Dash et al., 2009). Depending of which complexes are present in a cyanide solution, it can be classified as: i) free cyanide; when just cyanide ion and hydrogen cyanide are present, ii) weak-acid dissociable (WAD) cyanide; when, in addition to free cyanide, relatively weak complexes such as those with silver, cadmium, copper, mercury, nickel or zinc are present, iii) total cyanide; when free and WAD cyanide is present together with metals that form strong cyanide complexes, for example, iron, cobalt and gold (Angove & Acar, 2016).

2.1.2. Cyanide in industry

The capacity of the cyanide anion to form complexes of different stabilities with different metals has led to its use in different industries as a potent leachate (Veiga et al., 2014) as well as in metallurgic processes, electroplating, pesticide production, cosmetics, coal processing and synthetic fiber production (Mudder et al., 2004).

Gold processing and electroplating processes use approximately 20% of the approximately 1.1 million tons of cyanide produced annually worldwide (Mudder et al., 2004). In metallurgic industries, cyanide is used for leaching gold and silver from ores, making them soluble in alkaline conditions (Kuyucak & Akcil, 2013). This process allows recovery of up to 80 - 90 % of the metal (Veiga et al., 2014). In the electroplating industry, cyanide metal complexes in solution are used to generate a thin metal coating on an object to increase its anti-corrosion, abrasion resistance, and/or aesthetic qualities.

2.2.3. Cyanide toxicity

The capacity of cyanide to bind to metals also confers a high degree of toxicity (Hendry-Hofer et al., 2019; Leavesley et al., 2008). Almost all (if not all) the organisms known today use transition metal ions as cofactors of structural proteins and metabolic enzymes. One important example is the electron transport chain required for ATP production (Cooper & Brown, 2008). This electron transport chain is a multiprotein complex that transfers electrons between its subunits to a final acceptor (in the case of aerobic organisms the final acceptor is the oxygen). The capacity to transfer electrons between subunits in most of these cases is due to the presence of embedded metals in the protein structure (Crane et al., 1991). Cyanide binds to Fe^{+3} in heme-containing proteins, inhibiting the terminal cytochrome complex IV of the electron transport chain.

In multicellular organisms, such as humans, exposure to cyanide causes hard breathing, vomiting, headache, blood disorders, and higher doses causes damage in the brain, heart, kidneys and can produce coma and death (Hendry-Hofer et al., 2019). Sublethal doses of cyanide in the environment have negative effects in at least osmoregulation, early development, growth, swim, fat gain, and spermatogenesis (Eisler, 1991).

Due to these effects, governments have established cyanide limits in natural environments and industrial wastewater. EPA limit established is 0.2 mg/L of total cyanide in aquatic environments and 0.05 mg/L for drinking water (Environmental Protection Agency. 2010). Brazil has a 1 mg/L total cyanide limit for effluent discharge (Resolução CONAMA Nº 430 DE 13/05/2011).

2.2.4. Nitrilases

Nitrilases (EC: 3.5.5.1) are a superfamily of proteins characterized by an alpha-beta-beta-alpha (αββα) fold tertiary structure with the association of two monomers in a dimer as the basic catalytic unit (Sewell et al., 2003). The activity of these enzymes is conferred by a catalytic triad formed by a nucleophilic cysteine, a glutamic acid and a lysine and does not require any cofactor or prosthetic group (Brenner, 2002).

Pace & Brenner (2001) classified the members of this superfamily into thirteen branches with branch 1 corresponding to enzymes that hydrolyze the nitrile group into ammonia and its respective carboxylic acid with thyoimidate as an intermediate (Fig. 46). The dimers of these enzymes typically form large helical aggregates of several subunits (Thuku et al., 2009). For example, the cyanide dihydratase of *Bacillus pumilus* is reported to form an 18-subunit oligomer (Jandhyala et al., 2003); whereas the homolog in

Pseudomonas stutzeri forms a 14-subunit oligomer (Sewell et al., 2003). The nitrile substrates of these enzymes include aliphatic nitriles, aromatic nitriles, aryl-acetonitriles, among others (Black et al., 2015; Robertson et al., 2004).

In general, nitrilases are economically important in several biotechnology industries including the production of pharmaceutical intermediates, food additives and agrochemical precursors (Gong et al., 2012).

The regulation of nitrilase gene expression is not well understood. Some of them apparently are constitutively expressed while others are induced by nitriles (Chhiba-Govindjee et al., 2019). One well-characterized example is the transcriptional regulation of NitA from *Rhodococcus rhodochrous* J1. The *nitA* gene is part of an operon containing the gene for the transcriptional regulator NitR which in turn is activated by the amide gamma-caprolactam or by isovaleronitrile. Upon its activation, NitR induces the expression of both *nitA* and *nitR*, generating a positive feedback loop (Komeda et al., 1996).

2.2.5 Cyanide degrading-nitrilases

Two types of nitrilases can degrade cyanide through a hydrolytic pathway: cyanide hydratases (CHTs) and cyanide di-hydratases (CynDs). CHTs convert cyanide into formamide using one water molecule whereas CynDs convert it to formic acid and ammonia. The general reaction of CynD proceeds through the formation of a thioimidate intermediate as depicted in Figure 56.

At the structural level, it is accepted that CynD and CHT reactions occurs with the participation of the catalytic triad common for all nitrilases (described above; Pace & Brenner, 2001). Figure 57 shows a proposed mechanism of these reactions (Fernandes et al., 2006; Stolz et al., 2019). After the binding of the HCN substrate stabilized by the lysine and a water molecule (Fig. 57, step 1), cysteine attacks the carbon of HCN forming the thioimidate (step 2). Then, glutamate activates a water molecule that consequently attacks the thioimidate (step 3) generating a tetrahedral intermediate. This tetrahedral intermediate can follow two possible routes: i) in the case of CHTs, the unstable intermediate reorganizes by cleaving the S-C bond releasing formamide (Fig. 57, step 4 – 5 right), ii) in CynDs, the C-N bond is cleaved, releasing ammonia (step 4 left), followed by hydrolysis of S-C bond releasing, instead of formamide, formate (step 5 – 6 left). The structural basis for the predominance of one or the other pathway is still unknown.

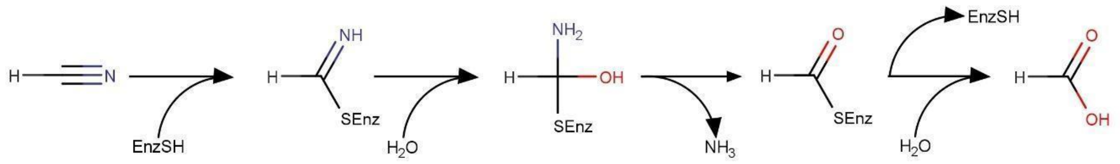


Figure 56. General reaction mechanism of CynD. CynD enzyme reacts with the substrate generating a thioimidate intermediate (step 1 – 2). Then, a water molecule attacks the thioimidate generating a tetrahedral intermediate (step 2 – 3). Rearrangements of the tetrahedral intermediate releases an ammonia molecule (step 3 – 4). Finally, a new rearrangement releases formate and reconstitute the enzyme (step 4 – 5).

At the moment, the ony well-characterized cyanide hydratases are derived from fungal genomes. The first enzyme described with this activity was from *Stemphylium loti* (Fry & Millar, 1972), followed by studies on the CHTs from other fungal species such as *Fusarium solani, Fusarium oxysporum, Micromonospora braunnam, Gloeocercospora* *sorghi, Leptosphaeria maculans*, and *Aspergillus niger* (Akinpelu et al., 2018; Dumestre et al., 1997; Ping Wang; Hans D. VanEtten, 1992; Rinágelová et al., 2014; Sexton & Howlett, 2000).

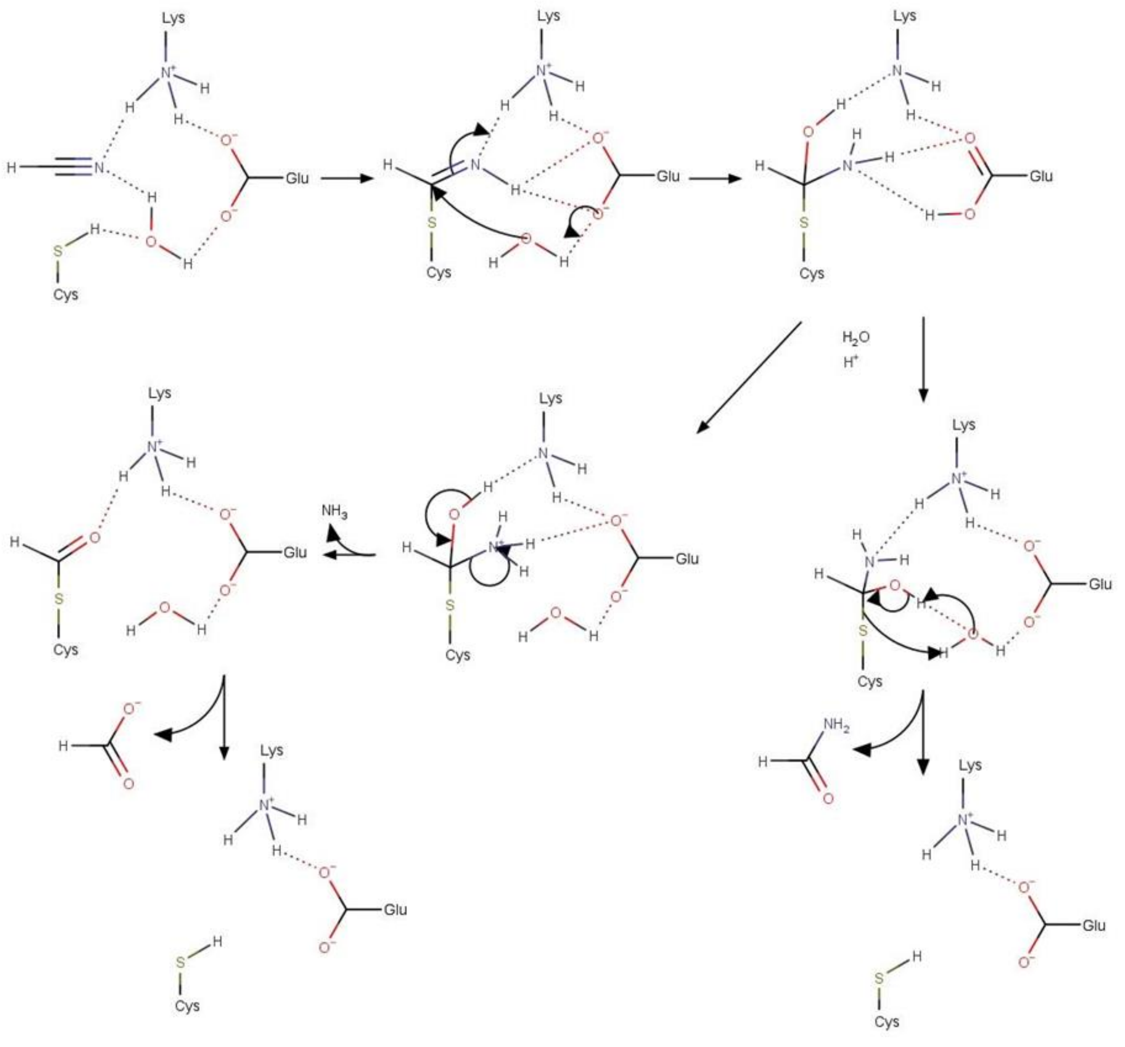


Figure 57. Proposed mechanism for CynD and CHT. Binding of cyanide occurs through hydrogen bond between lysine and a water molecule stabilized by glutamate (step 1). A nucleophilic attack of the sulfur atom to the carbon of cyanide occurs to form the thioimidate intermediate (step 1 to 2). Then, a water molecule activated by the glutamate attacks the carbon of the thioimidate (step 2) forming a tetrahedral intermediate (step 3). From here the reaction can follow two pathways: right) rearrangements of the tetrahedral intermediate involving a new water molecule release formamide and reconstitute the enzyme, left) a different rearrangement of the tetrahedral intermediate first release ammonia and then formate reconstituting the enzyme. Figure adapted from Fernandes et al., 2006 and Stolz et al., 2019.

On the other hand, only three CynDs have been experimentally characterized, all in bacterial species: CynD from *Bacillus pumilus, Pseudomonas stutzeri* and *Alcaligenes xylosoxidans* (Ingvorsen et al., 1991; Meyers et al., 1993; Watanabe et al., 1998).

2.2.6 Structure of nitrilases

Some proteins with nitrilase domains have been described structurally but none of them belong to the CynD or CHTs groups. The available structures with at least 30 % identity with CynD are NitA from *Pseudomonas fluorescens* (PDB ID: 6ZBY, identity: 38 %), Nit6803 from *Synechocystis sp*. (PDB ID: 3WUY, identity: 35 %), and Nit4 from *Arabidopsis thaliana* (PDB IDs: 6I5T, 6I5U, 6I00, identity: 30 %) (Figure 58).

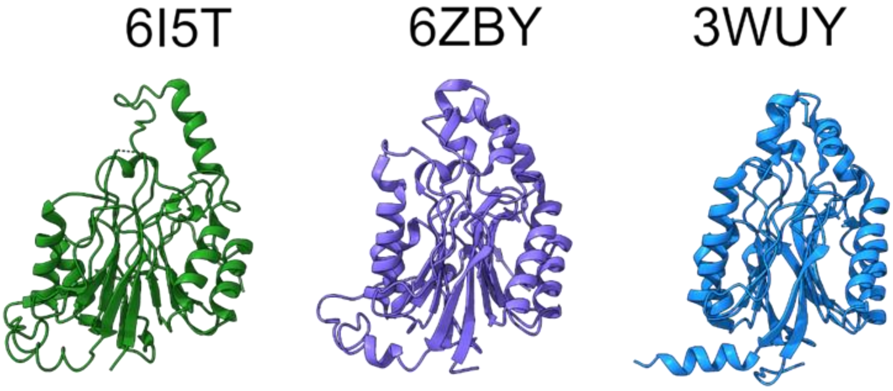


Figure 58. Closest CynD homologues. Nit4 from *Arabidopsis thaliana* (PDB ID: 6I5T) (left), NitA from *Pseudomonas fluorescens* (PDB ID: 6ZBY) (middle), and Nit6803 from *Synechocystis* (PDB ID: 3WUY) share 30, 38, and 35 % identity, respectively with CynD from *Bacillus pumilus*. These three proteins conserve the αββα core characteristic of nitrilases.

All of these three structures share the core αββα fold characteristic of nitrilases. It is therefore highly likely that CynD and CHT also have this core structural motif. However, finer features such as specific side chain positions and external loop conformation are not reliable from these homology models.

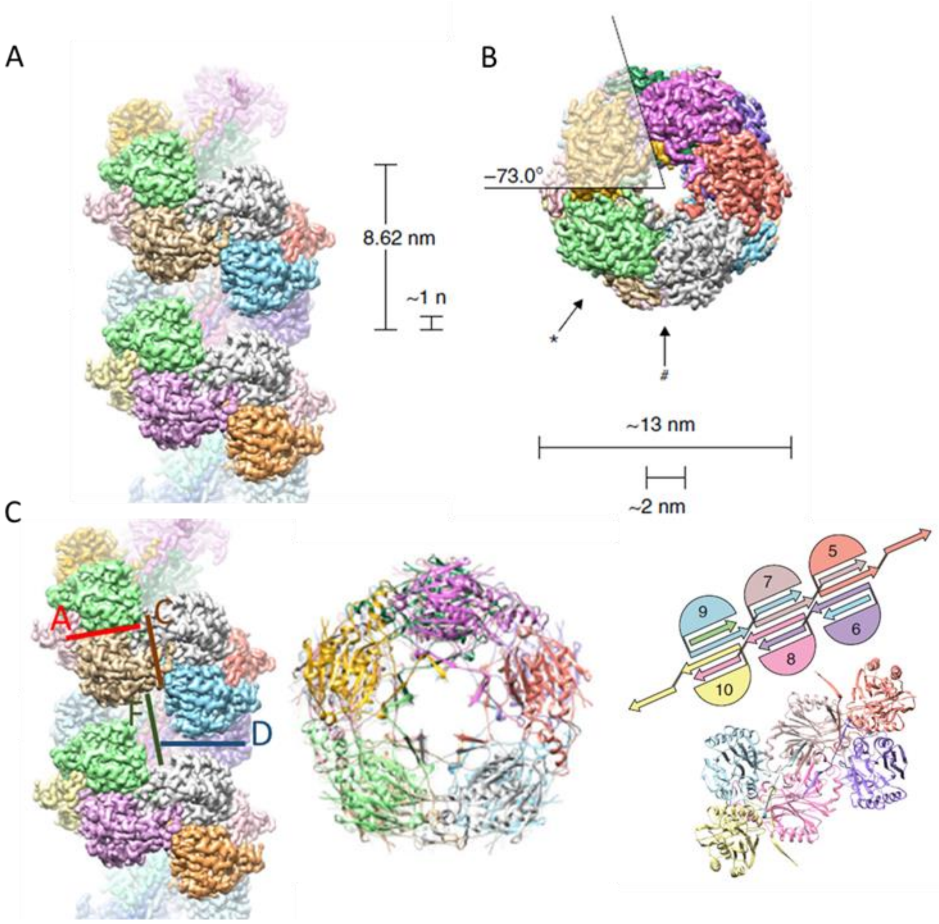


Figure 59. Structure of the spiral-forming nitrilase Nit4. A) Spiral structure is a left-handed helix with 4.9 dimers by turn and 8.62 nm by turn. B) It has a 13 nm diameter with a center hollow of 2 nm. C) (left) The spiral is stabilized by an interface between monomers (interface A) and another between dimers (interface C), interfaces F and D do not contribute to spiral stabilization. (middle and right) C-terminal regions forms beta-sheet that forms a crisscrossed structure that also contributes to spiral stabilization. Each C-terminal presents two beta-sheets that interacts with other three beta sheets from different monomers. Figure modified from Mulelu et al., 2019.

The only helical filament forming nitrilase with known high-resolution structure is Nit4 from *Arabidopsis thaliana* (3.4 Å resolution; PDB IDs: 6I5T, 6I5U, 6I00). As mentioned before, this protein shares only 30 % identity with CynD from *Bacillus pumilus*. The quarternary structure can be described as a left-handed helix with a rise of 8.62 nm per turn formed by 4.9 dimers (Figure 59A). The filament has a 13 nm diameter with a hollow center of 2 nm (Figure 59B). The spiral structure is stabilized by interface interactions between the dimers and crisscrossed beta-sheets in the center of the spiral (Figure 59C) (Mulelu et al., 2019).

2.2.- OBJECTIVES

The remainder of this chapter is presented in the form of a manuscript prepared for submission to a peer-reviewed journal.

The main objective of these studies was increasing our knowledge regarding cyanide metabolism in bacteria with the aim to generate improved bioremediation processes. The following specific objectives were persued during this study:

- Isolation of bacterial strains capable of degrading cyanide.
- Genomic sequencing of a *Bacillus safensis* strain with the capacity to degrade cyanide.
- Identification, cloning, expression, and purification of an enzyme responsible for cyanide degradation by the *Bacillus safensis* strain.
- Biochemical and bioinformatic analysis of the enzyme responsible to degrade cyanide in *Bacillus safensis*.

2.3.- PREPARED MANUSCRIPT

Prepared manuscript can also be found at:

https://www.biorxiv.org/content/10.1101/2021.11.27.470173v1

Isolation of a *Bacillus safensis* from mine tailings in Peru, genomic characterization, and characterization of its cyanide-degrading enzyme CynD

Santiago Justo Arevalo*[1,2,] Daniela Zapata Sifuentes[1,2,] Andrea Cuba Portocarrero[1], Michella Brescia Reategui[1], Claudia Monge Pimentel[1], Layla Farage Martins[2], Paulo Marques Pierry[2], Carlos Morais Piroupo[2], Alcides Guerra Santa Cruz[1], Mauro Quiñones Aguilar[1], Chuck Shaker Farah[2], João Carlos Setubal[2], Aline Maria da Silva[2]

1.- Facultad de Ciencias Biológicas, Universidad Ricardo Palma, Lima, Perú.

2.- Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil.

* Correspondence: santiago.jus.are@usp.br

2.3.1.- ABSTRACT

Cyanide is widely used in industry as a lixiviant due to its capacity to tightly bind metals. This property also imparts cyanide toxicity. Physical, chemical, and biological treatments have been used for cyanide remediation in industries; however, none of them meet the desired characteristics: efficiency, low-cost and low-environmental impact. A better understanding of metabolic pathways and biochemistry of enzymes involved in cyanide degradation is necessary to improve cyanide bioremediation. We have isolated three cyanide-degrading Bacillus from water in contact with mine tailings from Lima, Peru, and classified them as *Bacillus safensis* PER-URP-08, *Bacillus licheniformis* PER-URP-12, and *Bacillus subtilis* PER-URP-17 based on 16S rRNA gene sequencing and core genome analyses. Additionally, core genome analyses of 132 publicly available genomes of Bacillus pumilus group allowed us to reclassify some strains and identify two strains that did not match with any known species of the Bacillus pumilus group. We searched for cyanide-degradation pathways in the genomes of these three strains and identified putative B. licheniformis PER-URP-12 and B. subtilis PER-URP-17 rhodaneses and B. safensis PER-URP-08 cyanide dihydratase sequences possibly involved in cyanide degradation. Characteristic C-terminal residues differentiate CynD from B. pumilus and B. safensis, and, in contrast to CynD from B. pumilus C1, recombinant CynD from Bacillus safensis PER-URP-08 remains active up to pH 9. Moreover, transcripts of B. safensis PER-URP-08 CynD (CynDPER-URP-08) are strongly induced in the presence of cyanide. Our results warrant further investigation of B. safensis PER-URP-08 and CynDPER-URP-08 as potential tools for cyanide-bioremediation.

2.3.2.- INTRODUCTION

Cyanide is a highly toxic compound used in several industrial processes (Mudder et al., 2004) given its capacity to form tight complexes with different metals (Dash et al., 2009) (Hendry-Hofer et al., 2019; Leavesley et al., 2008). Industries that generate cyanide-containing wastes must reduce its concentration before release the environment, and as such proper strategies have to be implemented for cyanide remediation (Kuyucak &

Akcil, 2013). Cyanide bioremediation by bacteria that express nitrilases is one possible low-cost and environmental-friendly approach (Dash et al., 2009).

Nitrilases are a superfamily of proteins whose subunits share a common tertiary structure consisting of an alpha-beta-beta-alpha fold and a dimer as a basic unit. This superfamily has been divided into thirteen branches with branch one corresponding to enzymes that cleave the nitrile group into ammonia and its respective carboxylic acid. The other twelve branches are structurally similar, but their catalytic activity does not involve cleavage of nitriles (Pace & Brenner, 2001).

Two types of nitrilases can degrade cyanide through a hydrolytic pathway: cyanide hydratases (CHTs) and cyanide dihydratases (CynDs). CHTs convert cyanide into formamide, consuming one water molecule in the reaction, and are present in fungal genomes. The first enzyme described with this activity was from *Stemphylium loti* (Fry & Millar, 1972). Subsequently, CHTs from other fungal species were studied, including *Fusarium solani, Fusarium oxysporum, Gloeocercospora sorghi, Leptosphaeria maculans*, and *Aspergillus niger* (Akinpelu et al., 2018; Dumestre et al., 1997; Ping Wang; Hans D. VanEtten, 1992; Rinágelová et al., 2014; Sexton & Howlett, 2000). On the other hand, the only CynDs that have been experimentally studied are derived from bacterial species: *B. pumilus, P. stutzeri* and *Alcaligenes xylosoxidans* (Ingvorsen et al., 1991; Meyers et al., 1993; Watanabe et al., 1998). The reaction catalyzed by CynDs consumes two water molecules and generates formic acid and ammonia. Both, CynDs and CHTs, typically form large helical aggregates of several subunits (Thuku et al., 2009). For example, the CynD of *Bacillus pumilus* is reported to form an 18-subunit oligomer (Jandhyala et al., 2003) whereas the homolog in *Pseudomonas stutzeri* forms a 14-subunit oligomer (Sewell et al., 2003)

Several *Bacillus* species have been shown to be capable of metabolizing cyanide using different routes, for instance: Β-cyanoalanine synthase in Bacillus megaterium (Castric & Strobel, 1969), gamma-cyano-alpha-aminobutyric acid synthase in B. stearothermophilus (Omura et al., 2003), rhodanase in *Bacillus cereus* (Itakorode et al., 2019), and CynD in *Bacillus pumilus* (Meyers et al., 1993). On the other hand, some other cyanide-degrading *Bacillus* species have still unknown metabolic routes (Al-Badri et al., 2020; Javaheri Safa et al., 2017; Mekuto et al., 2014).

The *Bacillus pumilus* group consists mainly of three species: *Bacillus altitudinis, Bacillus safensis* and *Bacillus pumilus*. These three species share more than 99 % sequence identity in their 16S rRNA gene (Liu et al., 2013), hampering taxonomical classification based solely on this locus. Studies using multiple phylogenetic markers have demonstrated that ~50 % of the *Bacillus pumilus* group genomes deposited in NCBI database could be misclassified (Espariz et al., 2016).

It is plausible to speculate that CynDs isolated from *Bacillus* strains from diverse environments could present different properties, some of which could have certain properties better suited for certain industrial applications. Therefore, the characterization of CynD from other species can expand our understanding on the functioning and plasticity of this enzyme. Furthermore, some aspects of the biology of this enzyme have not been thoroughly studied. For instance, it is known that the oligomeric state of CynD is strongly pH-dependent (Jandhyala et al., 2003); however, the effect on oligomerization at pHs greater than 9 has not been reported. Also, it is unknown whether CynD is constitutively expressed in basal metabolism or is part of a specific physiological response, for instance, induced by the presence of cyanide.

Here, we describe the isolation of three indigenous *Bacillus* strains from mine tailing in Peru and their respective genome sequences. We selected a strain that was most efficient in cyanide degradation and investigated its phylogenetic relationship with other species of the *Bacillus pumilus* group. We identified a gene coding for a cyanide dihydratase (CynD) that is most likely the enzyme responsible for cyanide degradation in this selected strain. A recombinant CynD was expressed and purified, its catalytic parameters were determined, and the quaternary structure was studied at different pHs. We also demonstrated that CynD transcripts are strongly induced in the presence of cyanide.

2.3.3.- MATERIAL AND METHODS

2.3.3.1.- Isolation of cyanide-degrading strains

Water in contact with mine tailing was collected from the Casapalca river near Casapalca and La Oroya mines located in San Mateo de Huanchor (Latitude -11.4067 Longitude -

76.3361 at 4221 MASL). The sample was collected in 2 L sterile bottles and transported at 4 °C.

One hundred mL of the sample was added to an Erlenmeyer flask containing 20 mL of 21 g/L sodium carbonate, 9 g/L sodium bicarbonate, 5 g/L sodium chloride and 0.5 g/L potassium nitrate. Cultures were incubated for 12 h at 37 °C and after this time 1 mg/L final concentration of cyanide in the form of sodium cyanide was added. The cultures were incubated for another 24 h at 37 °C. Samples of the medium were streaked in petri dishes with nutrient agar (5 g/L peptone, 5 g/L yeast extract, 5 g/L sodium chloride and 1 % agar) and incubated at 37 °C for 24 h. Single colonies were isolated in nutrient broth (5 g/L peptone, 5 g/L yeast extract, 5 g/L sodium chloride) supplemented with 20 % glycerol and stored at -80 °C.

Strains stored at -80 °C were reactivated at 37 °C in nutrient agar by streaking a sample. One isolated colony was inoculated in fresh nutrient broth and incubated at 37 °C overnight at 100 g. Next, the optical density at 600 nm (OD_{600nm}) of the culture was adjusted to 0.8 and 1 mL was centrifuged at 6000 g for 3 min. The pellet was washed twice with 0.2 M Tris-HCl pH 8 and resuspended in 1 mL of 0.2 M Tris-HCl pH 8 supplemented with 0.2 M NaCN. After 2 h of incubation at 30 °C, the culture was centrifuged at 6000 g and 10 µL of the supernatant was taken and diluted in 90 µL of milliQ water. Then, 200 uL of 0.5 % picric acid in 0.25 M sodium carbonate was added and heated for 6 min at 100 °C (Williams & Edwards, 1980). Finally, absorbance at 520 nm was measured and compared to a standard curve of NaCN.

2.3.3.2.- Strain identification by 16S rRNA gene sequencing

To determine the bacterial genera and/or species of the isolated strains, we used a fresh culture in nutrient agar. Five colonies from these cultures were transferred to 50 µL of milliQ water and then heated to 100 °C for 3 min in a dry bath. The samples were centrifuged at 10 000 g for 5 min and the supernatant was used as a template to amplify a fragment that includes the V6, V7 and V8 variable regions of 16S rRNA gene. One µL of the template, 25 pmol of F_primer, 5´ GCACAAGCGGTGGAGCATGTGG 3´, and of the R_primer, 5´ GCCCGGGAACGTATTCACCG 3´, were mixed with 1x Taq buffer, 1.5 mmol of MgCl2, 0.2 mmol of each dNTP, and 1 U Taq DNA polymerase (ThermoFisher Scientific) in a final reaction of 25 µL. The amplification program was an initial denaturation at 94 °C for 5 min followed by 30 cycles at 94 °C for 45 sec, 50 °C for 45 sec, and 72 °C for 1 min, with a final extension of 10 min at 72 °C. Five µL of the final reaction was used as a template for the sequencing reaction. Sequencing reaction was done using Big Dye terminator v3.1 cycle sequencing kit (ThermoFisher Scientific) consisting of a 1x sequencing buffer, 25 pmol F_primer or R_primer and 2 µL of Big Dye in a final volume of 20 µL. The program used was an initial denaturation at 94 °C for 5 min, followed by 40 cycles at 94 °C for 30 sec, 50 °C for 30 sec, and 60 °C for 4 min. After the sequencing reaction, 80 µL of 70 % isopropanol was added and the reaction tube was centrifuged at 4000 g at 4 °C for 40 min. Then the supernatant was discarded, and the sample was resuspended in 20 µL of milliQ water and injected in an ABI PRISM 3130XL genetic analyzer (ThermoFisher Scientific) (Central Analítica – IQ – USP). The obtained sequences were used to perform BLASTn (Altschul et al., 1990) searches against the Genbank/NCBI database (Benson et al., 2013) to identify most similar sequences.

2.3.3.3.- Genome sequencing, assembling and annotation

Bacterial cultures were grown in 2xTY broth (tryptone 16 g/L, yeast extract 10 g/L, and NaCl 5 g/L) at 37 °C for 18 h at 200 rpm. Genomic DNA purification was performed using the Wizard Genomic DNA Purification Kit (Promega). DNA integrity was evaluated by 1 % agarose gel electrophoresis stained with SYBRSafe (Invitrogen) and by Bioanalyzer 2100 using Chips Agilent DNA 12000. DNA concentration and purity were estimated using a NanoDrop One/OneC Microvolume UV-Vis Spectrophotometer (ThermoFisher Scientific). Shotgun genomic library was prepared using the Nextera DNA Library Prep (Ilumina) with total DNA input of 20-35 ng. The resulting indexed DNA library was cleaned up with Agencourt AMPure XP beads (Beckman Coulter) and fragment size within the range of 200-700 bp were verified by running in the 2100 Bioanalyzer using Agilent High Sensitivity DNA chip. Fragment library quantification was performed with KAPA Library Quantification Kit. Genomic libraries prepared for each strain were pooled and subjected to a run using an Ilumina MiSeq Reagent Kit v2 (2 x 250 cycles) which generated ~38 million raw paired-end reads with >75% of bases with quality score > 30.

The genome of strain PER-URP-08 was assembled with Discovar (v. 52488) (Weisenfeld et al., 2014). The genomes of strains PER-URP-12 and PER-URP-17 were assembled with A5 (v. 20160825) (Coil et al., 2015). Both softwares have adapters trimming and read quality checking as part of their respective assembly processes. The tool Medusa (Bosi et al., 2015) was used to generate final genome scaffolds using three sets of five reference genomes, one for each of the genome assemblies (Table 12). The final genome assemblies were submitted to the IMG/M (Chen et al., 2021) and to the NCBI (Benson et al., 2013; Tatusova et al., 2016) for automatic annotation.

Table 12. Accession numbers of reference genomes used in the assembly process.

Strain	IMG code
B. licheniformis 5NAP23	2654587692
B. licheniformis VTM3R78	2623620452
B. licheniformis B4091	2728369131
B. licheniformis GB2	2654587725
B. licheniformis 19TX	2770939458
B. pumilus JRS3	2667527735
B. pumilus SAFR-032	640753007
B. pumilus TUAT1	2684623054
B. pumilus SH-B11	2687453109
B. pumilus RI06-95	2639762961
B. subtilis HM-66	2671180306
B. subtilis PCI 246	2597490117
B. subtilis B4067	2667527922
B. subtilis ATCC19217	2630968642
B. subtilis J22	2505679041

2.3.3.4.- Phylogenetic analyses and identification of nitrilases

Annotated genomes belonging to *Bacillus pumilus*, *Bacillus safensis* or *Bacillus altitudinis* species in the category of “Chromosome”, “Scaffold” or “Complete” were downloaded from the Genbank/NCBI (Benson et al., 2013). Using the software cd-hit (Fu et al., 2012;

Li & Godzik, 2006) we identified coding sequences that are not duplicated and present in all the genomes (core genes). A total of 1766 core genes with more than 80 % identity and at least 90 % coverage were used in the analysis. Core genes were aligned using MAFFT with the FFT-NS-2 algorithm (Katoh & Standley, 2013). The resulting alignments were concatenated and used to calculate a distance matrix based on identity using Biopython (Cock et al., 2009). Phylogenetic inference by maximum likelihood was done using the concatenated alignments as the input and IQ-TREE2 (Minh et al., 2020) with the evolution model GTR+F+R3, ultrafast bootstrap 1000 (Hoang et al., 2018), and 1000 initial trees.

IMG/M tools (Chen et al., 2021) were used to identify nitrilases genes in the annotated genomes. Genes encoding the CN_hydrolase domain (PFAM code PF00795) were selected and checked regarding the genomic context and the related literature.

2.3.3.5.- Analysis of CynD sequences from Bacillus pumilus group genomes

Protein sequence annotations from genomes belonging to *Bacillus pumilus, Bacillus safensis* or *Bacillus altitudinis* in the category of “Chromosome”, “Scaffold” or “Complete” were downloaded from GenBank/NCBI (Benson et al., 2013) and used to construct a local database. We ran a BLASTp search (Altschul et al., 1990) using the query sequence AAN77004.1 against the constructed local database, and sequences with more than 90 % identity and 100 % coverage were identified as CynD orthologs. These sequences were aligned using MAFFT with the L-INS-I algorithm (Katoh & Standley, 2013). The resulting alignment was used as an input for the phylogenetic inference by maximum likelihood using IQ-TREE2 (Minh et al., 2020) with the evolution model JTTDCMut+I (Kosiol & Goldman, 2005), ultrafast bootstrap 1000 (Hoang et al., 2018), 1000 initial trees and -allnni option.

2.3.3.6.- Cloning, expression and purification of CynD

The coding sequence for CynD was amplified from genomic DNA of strain PER-URP-08 using the primers (restriction sites appear in uppercase): F_CynD (5' tttCATATGatgacaagtatttacccgaagtttc 3'), and R_CynD (5' tttCTCGAGcactttttcttcaagcaaccc 3') and cloned in the NdeI and XhoI sites of the pET-28 expression vector. Then, this plasmid was used as a template to amplify the CynD coding sequence with a C-terminal 6x-His tag using the primers F_CynD and R_2_CynD (5' tttGAATTCagtggtggtggtggtggtg 3') and cloned in the NdeI and EcoRI sites of the pET-11 plasmid.

To express CynD protein with the C-terminal His tag, we used the *Escherichia coli* BL21(DE3) pLysS strain, induced by 0.3 mmol/L of Isopropyl ß-D-1-thiogalactopyranoside for 23 h at 18°C. The cells were lysed by sonication using a lysis buffer (100 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl, 50 mmol/L Imidazole) and the suspension was clarified by centrifugation (13000 g). The supernatant was loaded onto Ni-NTA affinity resin (His-trap chelating 5 mL column), washed with 10 volumes of lysis buffer, and eluted with a gradient of 50 - 500 mmol/L Imidazole in 20 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl.

The eluted fractions were further purified by size exclusion chromatography using a Superdex pg 200 16/600 column and 20 mmol/L Tris-HCl pH 8.0, 100 mmol/L NaCl as running buffer. The eluted fractions were examined for purity by SDS-PAGE and fractions containing pure protein were concentrated in Amicon Ultra-15 Centrifugal filter units.

2.3.3.7.- Enzymatic assays of recombinant CynD

For the determination of K_m and V_{max}, enzymatic activity of recombinant CynD was measured at pH 8.0 using the Ammonia Assay Kit (Sigma-Aldrich). A concentration of 500 nM of CynD was used in all reactions with the following cyanide concentrations: 0.39, 0.625, 0.78, 1.25, 1.56, 2.5, 3.125, 5, 6.25, 12.5, 25 mmol/L, with a final volume of 111 uL at 30 °C. Measurements were taken on a plate reader, at 340 nm every 20 sec. Calculations of ammonia concentrations were carried out according to the manufacturer´s description.

To determine the optimal pH for CynD activity, reagent solutions were prepared (40 mmol/L NaCN, 100 mmol/L NaCl and 200 mmol/L Tris-HCl or N-cyclohexyl-3-aminopropanesulfonic acid (CAPS) at pH 8, 9 or 10, 11, respectively). Then, we added 5 µL of CynD in 100 mmol/L NaCl, Tris-HCl pH 8 to 45 µL of the reagent solutions to obtain a final concentration of CynD of 0, 5, 10, 15, or 20 µM. The reactions were incubated for 10 min at 37 °C. After that, 100 µL of picric acid 5 mg/mL, 0.25 M Na_2CO_3 was added, and the reactions were incubated at 99 °C for 6 min. Next, 30 µL of this reaction was transferred to a 96-well plate and absorbance at 520 nm was recorded. Final cyanide concentration was estimated based on calibration curves with cyanide concentrations between 0 – 40 mmol/L.

2.3.3.8.- Size Exclusion Chromatography coupled to Multi-Angle Light Scattering (SEC-MALS)

SEC-MALS analysis was used to determine the oligomeric state of recombinant CynD. Molar mass analysis was done in 100 mmol/L NaCl and 20 mmol/L Tris-HCl or CAPS at pH 8, 9 or 10, 11, respectively. Protein samples (100 µL injection of 3.47 mg/mL (89.36 µM) CynD) were separated using a Superdex 200 increase 10/300 GL coupled to a MiniDAWN TREOS multi-angle light scattering system and an Optilab rEX refractive index detector. Data analysis was performed using the Astra Software package, version 7.1.1 (Wyatt Technology Corp.).

2.3.3.9.- Transmission electron microscopy (TEM)

Ultra-thin carbon layer on lacey carbon-coated copper grids were negatively charged by a glow discharge of 25 sec at 15 mA. Four microliters of purified recombinant CynD in 20 mmol/L Tris-HCl pH 8.0 and 100 mmol/L NaCl in different concentrations (3.25 mg/mL or 1.625 mg/mL) were placed in the negative charged carbon-coated copper grid for 1 minute. The grids were washed twice with MilliQ water and then stained with 2 % uranyl acetate for 30 secs before blotting and air drying. Electron micrographs images were obtained using a JEOL JEM 2100 transmission electron microscope equipped with a Gatan ORIUS CCD detector at the Institute of Chemistry of the University of Sao Paulo.

2.3.3.10.- RT-qPCR to evaluate in vivo induction of *cynD* by cyanide

Bacillus strains were grown in meat broth (meat extract 1 g/L, yeast extract 2 g/L, peptone 5 g/L, NaCl 5g/L, MnCl2 10 mg/L) during 12 h at 30 °C, 200 rpm. One mL of the culture was centrifuged at 500 xg for 1 min, the supernatant was transferred to a clean tube and this tube was centrifuged at 8 000 xg for 3 min. The pellet was resuspended in 1 mL of NaCN ([CN-] 100 ppm) in milliQ water. Controls were resuspended in 1 mL milliQ water without NaCN. The tubes were incubated without agitation at 30 °C for 4 h and 100 µL were retrieved to measure cyanide concentration by the picric acid method (Williams & Edwards, 1980). Nine hundred µL was centrifuged, and the bacterial pellet was used immediately for total RNA extraction.

Total RNA extraction was done using Trizol-chloroform protocol. Briefly, bacterial pellets were treated with 100 µL of lysozyme 3 mg/mL at 37 °C for 30 min, and extraction was done using a mixture of 5:1 trizol:chloroform. After the extraction, the phase containing RNA was separated and the RNA was precipitated using isopropanol. RNA pellet was washed twice with 75 % ethanol and finally resuspended in 20 µL of Tris 20 mmol/L-DEPC. Total RNA concentration and purity were estimated in a NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer (ThermoFisher Scientific) and the integrity was evaluated in a 2100 Bioanalyzer using an Agilent RNA 6000 Pico chip. After DNase treatment, the samples were subjected to PCR to verify the absence of DNA contamination. cDNA synthesis was performed with 1 µg of the RNA and Thermo Scientific H Minus First Strand cDNA Synthesis kit. cDNA synthesis was verified by PCR and electrophoresis.

Amplification efficiency of the primers used in the RT-qPCR were verified using 300 nM of each primer and a 2-fold dilution series of the cDNA to generate a standard curve composed of 4 concentrations as follows: 62.5, 31.25, 15.625, and 7.8125 ng/µL. Each dilution reaction was performed in triplicate using the Maxima SYBR Green/ROX qPCR Master Mix kit (ThermoFisher Scientific) following the manufacturer instructions in a QuantStudio 3 equipment (ThermoFisher Scientific). Primers for the normalizing gene *rpsJ* (F_rpsJ 5’ TGAAACGGCTAAGCGTTCTG 3’, R_rpsJ 5’ ACGCATCTCGAATTGCTCAC 3’), and for the nitrilases *cynD* (F_cynD 5’ TGCCCAAAATGAGCAGGTAC 3’, R_cynD 5’ AAATGTCTGTGTCGCGATGG 3’) and *ykrU* (F_ykrU 5’ TTGGTGCGATGATTTGCTAT 3’, R_ykrU 5’ GTGTCTCTGCTTGTGCCTGT 3’) were tested for efficiency. The amplification efficiency of the qPCR reaction was calculated through the slope of the cDNA curve obtained for each primer pair.

Since primer pairs have showed similar efficiency, (*ykrU* = 119.108 %, *cynD* = 108.385 %, *rpsJ* = 104.55 %), we performed each qPCR assay in technical triplicates using 15.625 ng/µL of cDNA and the kit Maxima SYBR Green/ROX qPCR Master Mix (ThermoFisher Scientific) in a QuantStudio 3 equipment (ThermoFisher Scientific). ΔΔCT values were calculated in absence or presence of cyanide for the nitrilase genes *ykrU* and *cynD* using *rpsJ* as the normalizing gene. Three biological replicates were performed.

2.3.4.- RESULTS AND DISCUSSION

2.3.4.1.- Three *Bacillus spp*. isolates with capacity of cyanide degradation

Several colonies were obtained after selective enrichment in cyanide containing media of water in contact with mine tailing from Casapalca river near Casapalca and La Oroya mines located in San Mateo de Huanchor, Lima - Peru. Twenty colonies were screened for the ability to degrade cyanide (Table 13) and three colonies with the greatest efficiency in cyanide degradation (isolates 8, 12, and 17) were selected for further studies (Table 13).

Table 13. Cyanide removal percentage of the twenty isolates from mine tailings in Peru.

Strain	Cyanide removal (%)
1	19.45
2	4.11
3	10.41
4	35.82
5	37.9
6	42.15
7	1.27
8	69.92
9	44.96
10	2.47
11	17.45

12	66.11
13	10.42
14	40.82
15	14.9
16	21.45
17	63.31
18	0.41
19	29.82
20	32.41

Sequencing of the V6, V7, and V8 variable regions of 16S rRNA gene of the three selected isolates and analysis by BLAST showed that they belong to the genus *Bacillus* (Table 14). Isolates 12 and 17 were identified as *Bacillus licheniformis* and *Bacillus subtilis*, respectively (Table 14) and were named *Bacillus licheniformis* PER-URP-12 and *Bacillus subtilis* PER-URP-17. Isolate 8 was classified as a member of the *Bacillus pumilus* group based on the 16S rRNA gene sequence (Table 14). However, it was not possible to discriminate among the different species in the *Bacillus pumilus* group (Liu et al., 2013) and as such this isolated was provisionally named *Bacillus sp.* PER-URP-08.

Table 14. BLAST best-hits of the partial 16S rRNA gene for each tested strain.

Accession ID	Description	Identity	E-value	Reference	Strain
MZ723096.1	Bacillus safensis strain LgS5 16S ribosomal RNA gene, partial sequence	100%	2,00E-166	Unpublished	8
MZ722995.1	Bacillus pumilus strain YG35 16S ribosomal RNA gene, partial sequence	100%	2,00E-166	Unpublished	

MZ720806.1	Bacillus australimaris strain EPB15 16S ribosomal RNA gene, partial sequence	100%	2,00E-166	Unpublished	
MZ720801.1	Bacillus safensis strain EPB9 16S ribosomal RNA gene, partial sequence	100%	2,00E-166	Unpublished	
MZ707643.1	Bacillus pumilus strain XY36 16S ribosomal RNA gene, partial sequence	100%	2,00E-166	Unpublished	
MT642946.1	Bacillus licheniformis strain IND706 16S ribosomal RNA gene, partial sequence	100%	0,00E+00	Unpublished	12
MT642944.1	Bacillus licheniformis strain AD242 16S ribosomal RNA gene, partial sequence	100%	0,00E+00	Unpublished	
MT495615.1	Bacillus licheniformis strain HO-A7 16S ribosomal RNA gene, partial sequence	100%	0,00E+00	Unpublished	
MT487704.1	Bacillus licheniformis strain MPF77 16S ribosomal RNA gene, partial sequence	100%	0,00E+00	Unpublished	
MT487699.1	Bacillus licheniformis strain MPF71 16S ribosomal RNA gene, partial sequence	100%	0,00E+00	Unpublished	
LR595019.1	uncultured bacterium partial 16S rRNA gene	99,72%	0,00E+00	Unpublished	17
LR595005.1	uncultured bacterium partial 16S rRNA gene	99,72%	0,00E+00	Unpublished	
LR594929.1	uncultured bacterium partial 16S rRNA gene	99,72%	0,00E+00	Unpublished	
MN231725.1	Bacillus sp. (in: Bacteria) strain PL12_OD 16S ribosomal RNA gene, partial sequence	99,44%	0,00E+00	Unpublished	
MK493753.1	Bacillus subtilis strain BM2349 16S ribosomal RNA gene, partial sequence	99,44%	0,00E+00	Unpublished	

The genomes of these three strains were then sequenced in order to obtain a more accurate taxonomical classification as well as to gain insights about possible routes of cyanide degradation in the three strains under study. Table 15 shows a summary of assembly and annotation metrics of these genomes.

Table 15. Summary of IMG/M annotations of the three *Bacillus* genomes.

Strain	*B. safensis* **PER-URP-08**	*B. licheniformis* **PER-URP-12**	*B. subtilis* **PER-URP-17**
Assembly			
Coverage	694x	555x	573x
Number of contigs	17	5	2
Contig N50*[1]	3013666	3385269	4075214
Contig L50*[2]	1	1	1
Total base pairs	3718369	4282823	4075214
Annotation			
Total coding base pairs	3305799	3775188	3619459
G+C Percentage	41,61	45,88	43,8
Total Number of Genes	3872	4558	4183
Total Number of Protein-coding Genes	3758	4417	4032
RNA genes	114	141	151
rRNA genes	9	14	11
5s rRNA	7	8	10
16s rRNA	2	4	1
23s rRNA	0	2	0
tRNA genes	75	78	83
Other RNA genes	30	49	57
Proteins with predicted function	3197	3642	3498
Proteins without predicted function	651	775	534
Enzymes-coding genes	1054	1176	1144
Chromosomal cassettes	287	341	292
Genes coding transmembrane proteins	1062	1212	1171

*[1] Length of the shortest contig that when contigs are in decreasing order exceeds the 50 % of total genome length.

*[2] The smallest number of contigs that adding their lengths gives 50 % or more of the total genome length.

2.3.4.2.- *Bacillus sp.* PER-URP-08 is classified as *Bacillus safensis* based on core-genome comparisons

We performed a genome-wide comparative analysis of *Bacillus sp.* PER-URP-08 with 132 genomes of species from the *Bacillus pumilus* group retrieved from the GenBank/NCBI database (Benson et al., 2013) and identified 1766 coding sequences present in all the genomes (core genes). An identity matrix based on an alignment of these core genes showed three well defined branches and two genomes that do not belong to any of these three branches (Fig. 60).

Branch 1 (Fig. 60, brown names) contains several strains already characterized as *Bacillus altitudinis* by different methods (for instance: BA06, ku-bf1, B-388 (X. Fu et al., 2021)) and also 4 strains (TUAT1, MTCB 6033, SH-B11 and C4) previously annotated as *Bacillus pumilus*. However, our analysis clearly demonstrates that these four strains belong to *Bacillus altitudinis* and therefore require reclassification (Table 16) as previously suggested (Espariz et al., 2016; X. Fu et al., 2021). The core genes within the *Bacillus altitudinis* branch share more than 98 % identity whereas they share less than 89.5 % identity with core genomes of the other two branches (Fig. 61A).

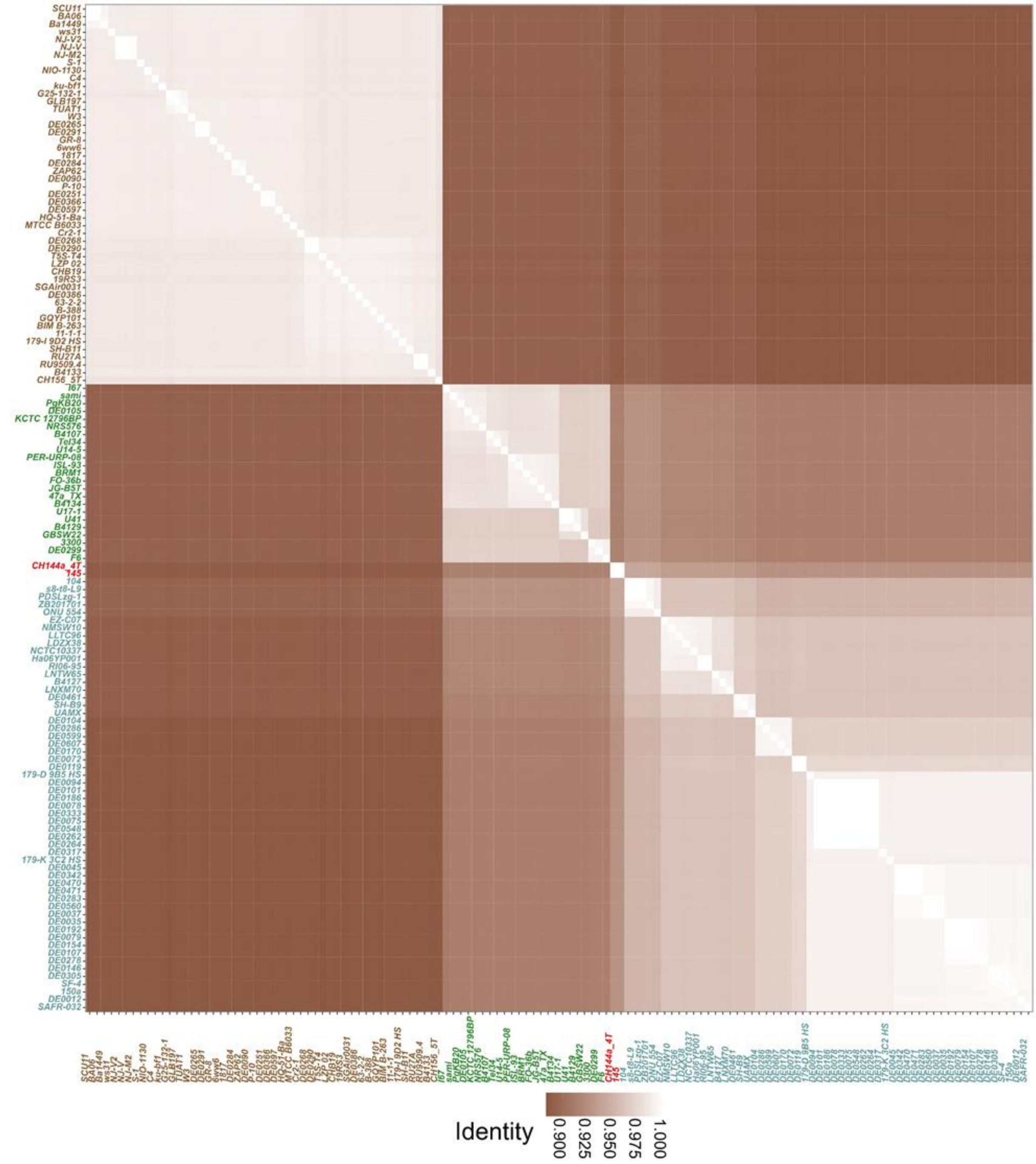


Figure 60. Core genome identity matrix to classified genomes of *Bacillus pumilus* group genomes. A) An identity matrix of 132 core genomes of *Bacillus pumilus* group showing delimitations between three species: *Bacillus altitudinis* (brown names), *Bacillus safensis* (green names), *Bacillus pumilus* (blue names). Two core genomes (red names) appear outside of these three species.

Identity of core genes in branch 2 is greater than 96 %, and this branch is more related to branch 3 (*Bacillus pumilus*, see below) than to branch 1 (*Bacillus altitudinis*) (Fig. 61B). Branch 2 (Fig. 60, green names) contains the *Bacillus safensis* type strain FO-36b (Satomi et al., 2006) as well as other strains already classified as *Bacillus safensis* such as B4107, B4134, and B4129 (Espariz et al., 2016). *Bacillus sp.* PER-URP-08 appeared inside this branch very near to the type strain FO-36b (99.2 % identity) (Fig. 50A) and so will be named *Bacillus safensis* PER-URP-08 from here on.

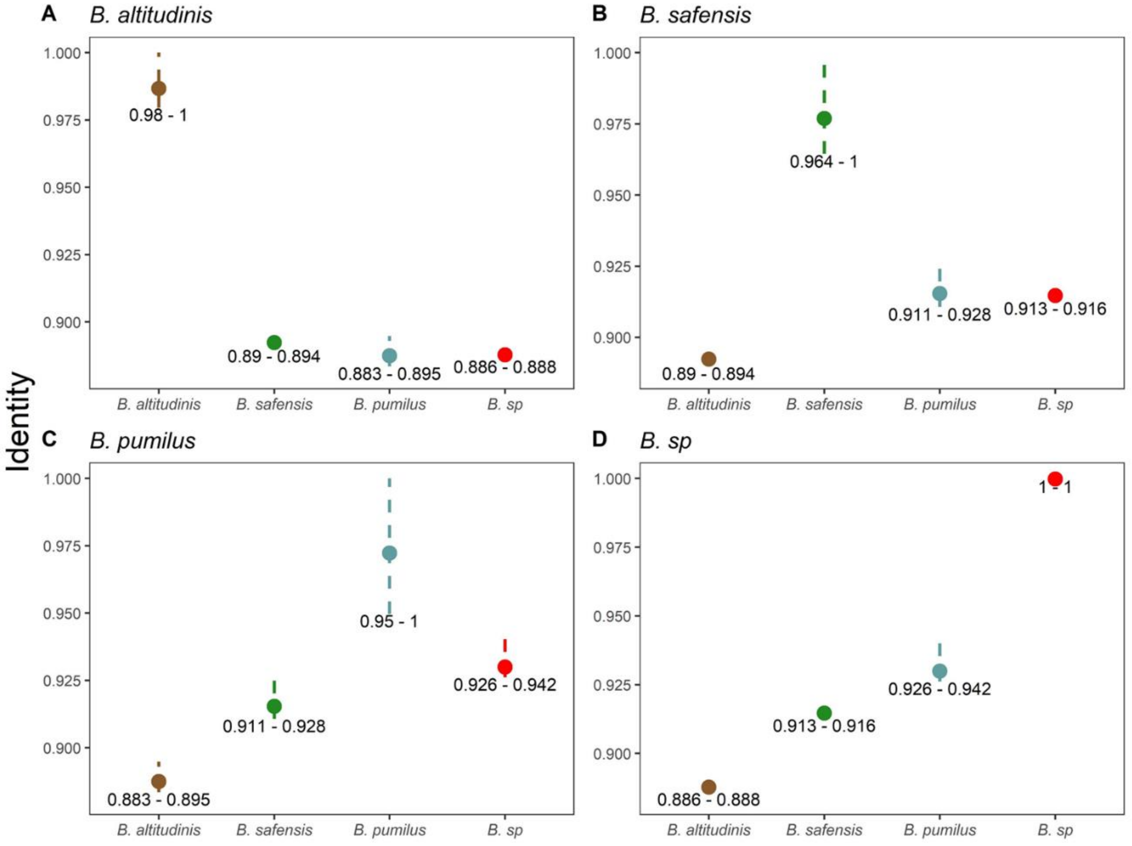


Figure 61. Ranges of genome identity between species of the *Bacillus pumilus* group. A – D) Plots showing the range of identity when compare *B. altitudinis* (B), *B. safensis* (C), *B. pumilus* (D) or *B. sp* (E) with itself or with other groups.

Branch 3 (Fig. 60, blue names) contains the SAFR-032 strain that was the first completely sequenced genome of *Bacillus pumilus* (Gioia et al., 2007; Stepanov et al., 2016). This branch 3 appears to be more heterogeneous than the other two branches (*Bacillus altitudinis* and *Bacillus safensis*) with more than 0.95 identity of the core genes of this branch (Fig. 61C).

Additionally, two genomes isolated from Mexico (CH144a_4T and 145) share less than 95 % identity with the branch 3 (between 92.6 and 94.2 % identity) and even less with branches 2 and 1 (between 91.3 and 91.6 % identity for branch 2 and between 88.6 and 88.8 % identity for branch 1) (Fig. 61D). The fact that these two genomes share less than 0.95 identity with all the other genomes in the analysis (Fig. 61D) indicates that CH144a_4T and 145 strains should be classified as different species outside the Bacillus pumilus group.

Table 16. Summary information of the 132 genomes used in the core genomes analysis.

Specie	GB_Specie	Strain	CynD_presence	Location	Assembly_ID
B. altitudinis	B. altitudinis	11-1-1	no	Belarus	GCA_013283915.1
	B. altitudinis	179-I 9D2 HS	no	USA	GCA_019037255.1
	B. altitudinis	1817	no	China	GCA_017161205.1
	B. altitudinis	19RS3	no	Argentina	GCA_013391605.1
	B. altitudinis	63-2-2	no	Belarus	GCA_016807685.1
	B. altitudinis	6ww6	no	China	GCA_017948365.1
	B. altitudinis	B-388	no	USA	GCA_000789425.2
	B. altitudinis	B4133	no	Netherlands	GCA_000828455.1
	B. altitudinis	BA06	no	China	GCA_000299555.2
	B. altitudinis	Ba1449	no	China	GCA_015689015.1
	B. altitudinis	BIM B-263	no	Belarus	GCA_015160895.1
	B. pumilus	C4	no	Egypt	GCA_001687085.1

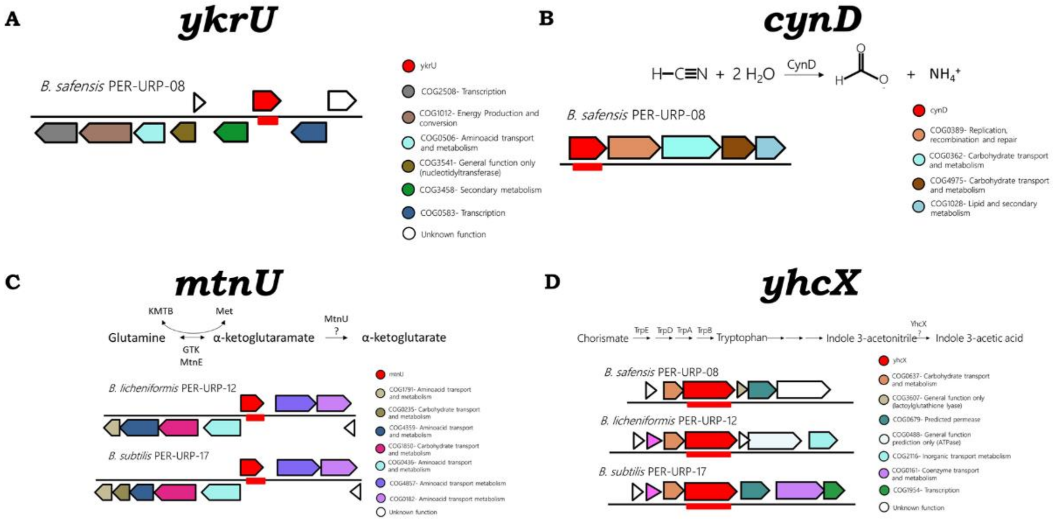
	B. altitudinis	CH156_5T	no	Mexico	GCA_008180475.1
	B. altitudinis	CHB19	no	Malaysia	GCA_004563755.2
	B. altitudinis	Cr2-1	no	China	GCA_007923025.1
	B. altitudinis	DE0090	no	USA	GCA_007682105.1
	B. altitudinis	DE0251	no	USA	GCA_008764185.1
	B. altitudinis	DE0265	no	USA	GCA_007681425.1
	B. altitudinis	DE0268	no	USA	GCA_007681435.1
	B. altitudinis	DE0284	no	USA	GCA_007681345.1
	B. altitudinis	DE0290	no	USA	GCA_007681315.1
	B. altitudinis	DE0291	no	USA	GCA_007681245.1
	B. altitudinis	DE0366	no	USA	GCA_007676515.1
	B. altitudinis	DE0386	no	USA	GCA_007676435.1
	B. altitudinis	DE0597	no	USA	GCA_007671735.1
	B. altitudinis	G25-132-1	no	China	GCA_015846075.1
	B. altitudinis	GLB197	no	China	GCA_001908475.1
	B. altitudinis	GQYP101	no	China	GCA_005849435.1
	B. altitudinis	GR-8	no	China	GCA_001191605.1
	B. altitudinis	HQ-51-Ba	no	not_collected	GCA_006007905.1
	B. altitudinis	ku-bf1	no	India	GCA_001543165.1
	B. altitudinis	LZP 02	no	China	GCA_019164215.1
	B. pumilus	MTCC B6033	no	India	GCA_000590455.1
	B. altitudinis	NIO-1130	no	India	GCA_001457015.1
	B. altitudinis	NIO-1130	no	not_collected	GCA_900094985.1
	B. altitudinis	NJ-M2	no	China	GCA_001431145.1
	B. altitudinis	NJ-V	no	China	GCA_001700735.1
	B. altitudinis	NJ-V2	no	China	GCA_001431785.1
	B. altitudinis	P-10	no	Indonesia	GCA_002741745.1
	B. altitudinis	RU27A	no	not_collected	GCA_900188195.1
	B. altitudinis	RU9509.4	no	not_collected	GCA_900119345.1
	B. altitudinis	S-1	no	not_collected	GCA_000225935.1
	B. altitudinis	SCU11	no	China	GCA_013307105.1
	B. altitudinis	SCU11	no	China	GCA_019355135.1
	B. altitudinis	SGAir0031	no	Singapore	GCA_002443015.2
	B. pumilus	SH-B11	no	Netherlands	GCA_001578165.1
	B. altitudinis	T5S-T4	no	Argentina	GCA_013391615.1
	B. pumilus	TUAT1	no	Japan	GCA_001548215.1
	B. altitudinis	W3	no	China	GCA_000972685.1
	B. altitudinis	ws31	no	China	GCA_016767855.1
	B. altitudinis	ZAP62	no	Mexico	GCA_011067205.1
B. pumilus	B. pumilus	104	no	USA	GCA_003034105.1
	B. pumilus	DE0104	no	USA	GCA_007679395.1
	B. pumilus	DE0170	no	USA	GCA_007678395.1
	B. pumilus	DE0286	no	USA	GCA_007676815.1
	B. pumilus	DE0599	no	USA	GCA_007665445.1
	B. pumilus	DE0607	no	USA	GCA_007665325.1
	B. pumilus	Ha06YP001	no	USA	GCA_003020795.1

	B. pumilus	ONU 554	no	Ukraine	GCA_014489355.1
	B. pumilus	PDSLzg-1	no	China	GCA_001704975.1
	B. pumilus	RI06-95	no	USA	GCA_001183525.1
	B. pumilus	s8-t8-L9	no	Atlantic Ocean	GCA_018128785.1
	B. pumilus	ZB201701	no	China	GCA_004006455.1
	B. pumilus	150a	yes	Mexico	GCA_003571425.1
	B. pumilus	179-D 9B5 HS	yes	USA	GCA_019037765.1
	B. pumilus	179-K 3C2 HS	yes	USA	GCA_019036865.1
	B. pumilus	B4127	yes	Netherlands	GCA_000828345.1
	B. pumilus	DE0012	yes	USA	GCA_007680695.1
	B. pumilus	DE0035	yes	USA	GCA_007680335.1
	B. pumilus	DE0037	yes	USA	GCA_007680315.1
	B. pumilus	DE0045	yes	USA	GCA_007680195.1
	B. pumilus	DE0072	yes	USA	GCA_007679805.1
	B. pumilus	DE0075	yes	USA	GCA_007679755.1
	B. pumilus	DE0078	yes	USA	GCA_007679665.1
	B. pumilus	DE0079	yes	USA	GCA_007679675.1
	B. pumilus	DE0094	yes	USA	GCA_007679515.1
	B. pumilus	DE0101	yes	USA	GCA_007679485.1
	B. pumilus	DE0107	yes	USA	GCA_007679415.1
	B. pumilus	DE0119	yes	USA	GCA_007679215.1
	B. pumilus	DE0146	yes	USA	GCA_007678815.1
	B. pumilus	DE0154	yes	USA	GCA_007678705.1
	B. pumilus	DE0186	yes	USA	GCA_007678135.1
	B. pumilus	DE0192	yes	USA	GCA_007678035.1
	B. pumilus	DE0262	yes	USA	GCA_007677155.1
	B. pumilus	DE0264	yes	USA	GCA_007677125.1
	B. pumilus	DE0278	yes	USA	GCA_007676935.1
	B. pumilus	DE0283	yes	USA	GCA_007676865.1
	B. pumilus	DE0305	yes	USA	GCA_007674165.1
	B. pumilus	DE0317	yes	USA	GCA_007674015.1
	B. pumilus	DE0333	yes	USA	GCA_007673765.1
	B. pumilus	DE0342	yes	USA	GCA_007673705.1
	B. pumilus	DE0461	yes	USA	GCA_007667685.1
	B. pumilus	DE0470	yes	USA	GCA_007667505.1
	B. pumilus	DE0471	yes	USA	GCA_007667475.1
	B. pumilus	DE0548	yes	USA	GCA_007666215.1
	B. pumilus	DE0560	yes	USA	GCA_007666085.1
	B. pumilus	EZ-C07	yes	Russia	GCA_003301255.1
	B. pumilus	LDZX38	yes	China	GCA_002998475.1
	B. pumilus	LLTC96	yes	China	GCA_002998365.1
	B. pumilus	LNTW65	yes	China	GCA_002998415.1
	B. pumilus	LNXM70	yes	China	GCA_002998395.1
	B. pumilus	NCTC10337	yes	not_collected	GCA_900186955.1
	B. pumilus	NMSW10	yes	China	GCA_002998335.1
	B. pumilus	SAFR-032	yes	not_collected	GCA_000017885.4

	B. pumilus	SF-4	yes	Pakistan	GCA_009937765.1
	B. pumilus	SH-B9	yes	Netherlands	GCA_001578205.1
	B. pumilus	UAMX	yes	Mexico	GCA_013423765.1
B. safensis	B. safensis	B4107	no	Netherlands	GCA_000828395.1
	B. safensis	NRS576	no	not_collected	GCA_900573445.1
	B. safensis	Tel34	no	Greece	GCA_016767355.1
	B. safensis	U14-5	no	Antarctica	GCA_001938665.1
	B. safensis	3300	yes	USA	GCA_007829795.1
	B. safensis	47a_TX	yes	USA	GCA_003610615.1
	B. safensis	B4129	yes	Netherlands	GCA_000828375.1
	B. safensis	B4134	yes	Netherlands	GCA_000828425.1
	B. safensis	BRM1	yes	Brazil	GCA_002077215.1
	B. safensis	DE0105	yes	USA	GCA_008764375.1
	B. safensis	DE0299	yes	USA	GCA_007674245.1
	B. safensis	F6	yes	Belarus	GCA_016803835.1
	B. safensis	FO-36b	yes	USA	GCA_003097715.1
	B. safensis	GBSW22	yes	China	GCA_002998315.1
	B. safensis	I67	yes	Brazil	GCA_012972765.1
	B. safensis	ISL-93	yes	Chile	GCA_018614995.1
	B. safensis	JG-B5T	yes	Germany	GCA_003284765.1
	B. safensis	KCTC 12796BP	yes	South Korea	GCA_001895885.1
	B. pumilus	PER-URP-08	yes	Peru	GCA_016629615.1
	B. safensis	PgKB20	yes	South Korea	GCA_008244765.1
	B. safensis	sami	yes	Pakistan	GCA_003660145.1
	B. safensis	U17-1	yes	Antarctica	GCA_001938705.1
	B. safensis	U41	yes	Antarctica	GCA_001938685.1
B. sp	B. pumilus	145	no	Mexico	GCA_003431975.1
	B. pumilus	CH144a_4T	no	Mexico	GCA_008180455.1

2.3.4.3.- A cyanide dihydratase is likely the responsible for cyanide degradation in B. safensis PER-URP-08

To gain insight regarding the enzymes responsible for cyanide metabolism in the strains *B. safensis* PER-URP-08, *B. licheniformis* PER-URP-12, and *B. subtilis* PER-URP-17, we first searched for genes coding for proteins related to nitrilases. The PFAM database annotates homologs of nitrilases as CN_hydrolases under the PFAM code PF00795. Using IMG/M system tools (Chen et al., 2021), we determined the presence of three, two, and two proteins containing CN_hydrolase domains in *B. safensis* PER-URP-08, *B. licheniformis* PER-URP-12, and *B. subtilis* PER-URP-17, respectively (Fig. 62). Both *B. licheniformis* PER-URP-12 and *B. subtilis* PER-URP-17 present the genes *yhcX* (NCBI locus tags: EGI08_RS06285 and EGI09_16505, respectively) and *mtnU* (EGI08_RS08970 and EGI09_01680, respectively). YhcX is probably involved in the degradation of indole-3-acetonitrile, a sub product of tryptophan metabolism (Idris et al., 2007) (Fig. 62).



A
ykrU
B. safensis PER-URP-08
ykrU
COG2508- Transcription
COG1012- Energy Production and conversion
COG0506- Aminoacid transport and metabolism
COG3541- General function only (nucleotidyltransferase)
COG3458- Secondary metabolism
COG0583- Transcription
Unknown function
B
cynD
H—C≡N + 2 H2O
CynD
H
O
O
+ NH4+
B. safensis PER-URP-08
cynD
COG0389- Replication, recombination and repair
COG0362- Carbohydrate transport and metabolism
COG4975- Carbohydrate transport and metabolism
COG1028- Lipid and secondary metabolism
C
mtnU
KMTB
Met
Glutamine
α-ketoglutaramate
GTK
MtnE
MtnU
?
α-ketoglutarate
B. licheniformis PER-URP-12
B. subtilis PER-URP-17
mtnU
COG1791- Aminoacid transport and metabolism
COG0235- Carbohydrate transport and metabolism
COG4359- Aminoacid transport and metabolism
COG1850- Carbohydrate transport and metabolism
COG0436- Aminoacid transport and metabolism
COG4857- Aminoacid transport metabolism
COG0182- Aminoacid transport metabolism
Unknown function
D
yhcX
Chorismate
TrpE
TrpD
TrpA
TrpB
Tryptophan
Indole 3-acetonitrile
YhcX
?
Indole 3-acetic acid
B. safensis PER-URP-08
B. licheniformis PER-URP-12
B. subtilis PER-URP-17
yhcX
COG0637- Carbohydrate transport and metabolism
COG3607- General function only (lactoylglutathione lyase)
COG0679- Predicted permease
COG0488- General function prediction only (ATPase)
COG2116- Inorganic transport metabolism
COG0161- Coenzyme transport and metabolism
COG1954- Transcription
Unknown function

Figure 62. Proteins containing CN_hydrolase domain in the three genomes studied. Four CN_hydrolases containing-proteins were identified in the analyzed genomes. YkrU and CynD are present only in *B. safensis* PER-URP-08. MtnU was found in *B. licheniformis* PER-URP-12 and *Bacillus subtilis* PER-URP-17. YhcX was found in the three genomes.

On the other hand, MtnU has been described as a possible enzyme catalyzing the conversion of alpha-ketoglutaramate to alpha-ketoglutarate involved in the metabolism of methionine (Ellens et al., 2015; Sekowska & Danchin, 2002) (Fig. 62). None of the enzymes with a CN_hydrolase domain in *B. licheniformis* PER-URP-12 and *B. subtilis* PER-URP-17 appears to be responsible for cyanide degradation.

Apart from these proteins, Bacillus and other genera present proteins with rhodanese domains (PFAM codes PF12368 and PF00581) (Table 17) that are able to convert thiosulphate and cyanide to sulphite and thiocyanate (Cipollone et al., 2006; Itakorode et al., 2019). Also, it was shown that the capacity to detoxify cyanide by *Bacillus stearothermophilus* is increased in mutants showing higher rhodanese activity (Atkinson, 1975), and RdhA from *Pseudomonas aeruginosa* overexpressed in *E. coli* has been shown to provide protection against cyanide when overexpressed in *E. coli* (Cipollone et al., 2006). However, cyanide detoxification is not the only function described for rhodanases in prokaryotes, proteins with thiosulfate:cyanide sulfurtransferase has been involved in other possible functions: 1) The phage-shock protein E (PspE) from *E. coli* (Adams et al., 2002) and the shock protein Q9KN65 from *Vibrio cholerae* has been associated with other specific stress conditions (Heidelberg et al., 2000); 2) in *Acidiothiobacillus ferroxidans*, the proteins P15 and P16.2 are possibly involved in sulfur oxidation (Acosta et al., 2005). Furthermore, proteins containing rhodanese domains can have activities different to thiosulfate:cyanide sulfurtransferase: RdlA from *Halanaerobium congolense* is able to catalyze the reductive cleavage (Ravot et al., 2005) of thiosulfate or RdhA from *Azotobacter vinelandii* is part of a delivery system of selenodiglutathione (Melino et al., 2003).

Finally, several rhodanese domain containing proteins are present in the genomes analyzed here and in other bacterial genomes (Cipollone et al., 2007). The coexistence of several proteins with the same domain in the same organism suggests that different physiological roles could be covered for each of these proteins. If one or more of these rhodanese domain-containing proteins are responsible for the degradation of cyanide by B. licheniformis PER-URP-12 and B. subtilis PER-URP-17 needs further studies to be tested.

Table 17. Rhodanese domain coding ORFs in the three sequenced genomes.

NCBI Locus Tag	Function ID	Strain
EGI09_06665	pfam00581	*Bacillus subtilis* PER-URP-17
EGI09_09165	pfam00581	*Bacillus subtilis* PER-URP-17
EGI09_17495	pfam12368	*Bacillus subtilis* PER-URP-17
EGI09_17495	pfam00581	*Bacillus subtilis* PER-URP-17
EGI08_04135	pfam12368	*Bacillus licheniformis* PER-URP-12
EGI08_04135	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_15285	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_16030	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_16065	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_16070	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_17900	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI08_19145	pfam00581	*Bacillus licheniformis* PER-URP-12
EGI07_08430	pfam00581	*Bacillus safensis* PER-URP-08
EGI07_10975	pfam00581	*Bacillus safensis* PER-URP-09
EGI07_18045	pfam12368	*Bacillus safensis* PER-URP-10
EGI07_18045	pfam00581	*Bacillus safensis* PER-URP-11
EGI07_18105	pfam00581	*Bacillus safensis* PER-URP-12
EGI07_18110	pfam00581	*Bacillus safensis* PER-URP-13
EGI07_18115	pfam00581	*Bacillus safensis* PER-URP-14

B. safensis PER-URP-08 presents *yhcX* (EGI07_01665) but not *mtnU*. In addition, this strain carries two other proteins containing a CN_hydrolase domain, EGI07_17510 and

CynD (EGI07_08135). EGI07_17510 is a protein of unknown function whereas CynD homologs (Fig. 63) hydrolyzes cyanide to produce ammonia and formic acid (Dash et al., 2009; Ibrahim et al., 2015). We therefore carried out a series of experiments to test the hypothesis that CynD is the enzyme responsible for cyanide degradation in *B. safensis* PER-URP-08.

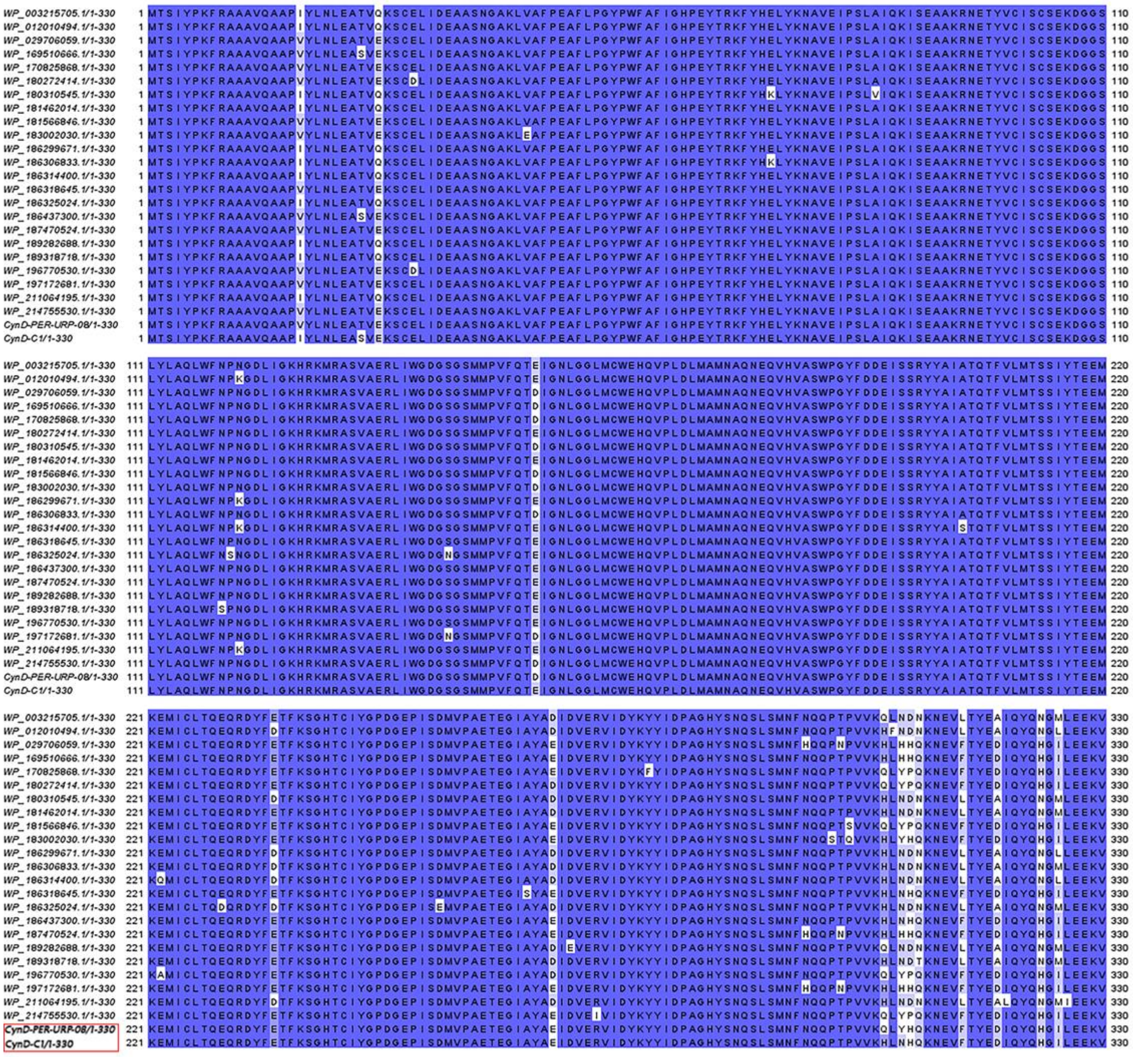


Figure 63. Alignments of identical protein group CynDs with CynD$_{\text{PER-URP-08}}$ and CynD$_{\text{C1}}$. Homologies of CynD from *Bacillus safensis* PER-URP-08 is clearly showed in the protein sequence alignments of several CynD homologs including those with tested enzymatic activity as CynD from *B. pumilus* strain C1.

2.3.4.4.- C-terminal residues differentiate CynD from *B. pumilus* and *B. safensis*

We first constructed a maximum likelihood (ML) phylogenetic tree based on the 132 core genomes of strains from *Bacillus pumilus* group (Fig. 64A) and searched for orthologs of CynD in the strains present in the ML tree (see Methods for details of the search).

The ML tree confirmed the three branches identified above (Fig. 60) and that two genomes (CH144a_4T and 145) do not belong to any of these branches (Fig. 64A). Intriguingly, CynD-encoding sequences were found in some representatives of *B. pumilus* (44 out of 56) and *B. safensis* (19 out of 23) but not in *B. altitudinis*. Three monophyletic *B. pumilus* and one monophyletic *B. safensis* clades lack CynD (Fig. 64A). This could be due to processes of gene gain and/or loss in the strains, and further studies are necessary to distinguish between these or other possibilities. It is also possible that some cynD genes were not sequenced in some genomes that are not completely closed.

Next, we identified twenty-three different sequences of CynD in the 132 genomes (Table 18) and a ML phylogenetic tree based on aminoacid sequences was constructed, including the sequences of the CynD from strain C1 (CynD$_{C1}$) (accession id: AAN77004.1) and of the CynD from *B. safensis* PER-URP-08 (CynD$_{PER\text{-}URP\text{-}08}$). A clear separation between CynD from *B. safensis* and from *B. pumilus* could be observed in the ML tree (Fig. 64B). Interestingly, CynD$_{C1}$ appear more related to the *B. safensis* group (Fig. 64B).

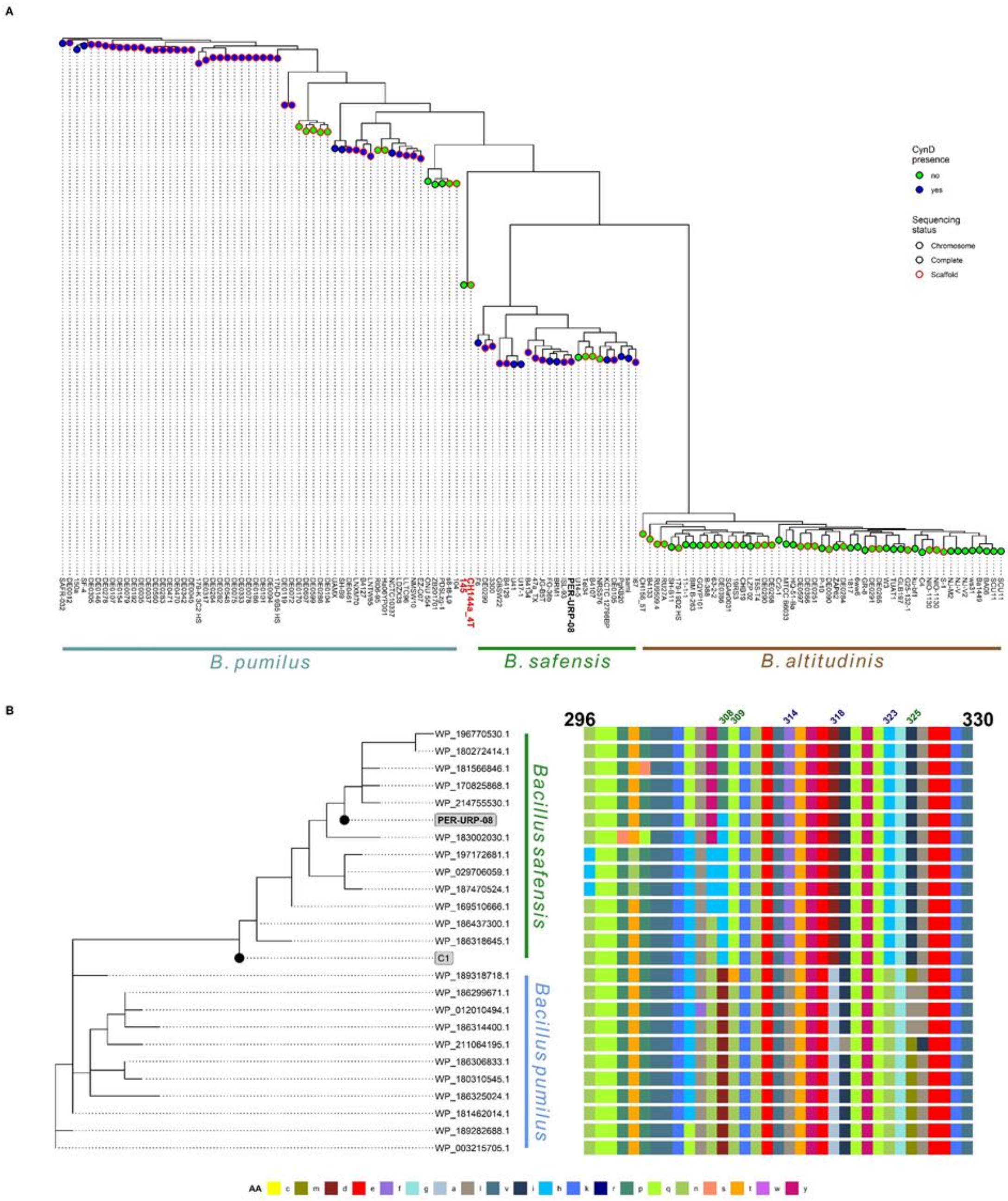


Figure 64. CynD is present in some genomes of *B. pumilus* and *B. safensis* and they are mainly differentiated by C-terminal residues. A) Maximum likelihood tree of core genomes of 132 *Bacillus pumilus* group strains showing separation between three species. Color of the circles represent absence (green) or presence (blue) of CynD homologue in the genome. Circles with black and red borders represent complete genomes ("chromosome" or "complete" sequencing status in NCBI) and possibly not complete genomes ("scaffold" sequencing status in NCBI). B) Maximum likelihood tree of full-length CynD sequences associated to an alignment of their C-terminal region (residues 296 to 330). Showed in number blue or green are the positions that are completely conserved in *Bacillus safensis* or *B. pumilus*, respectively.

Due to the several taxonomic misclassifications of strains belonging to the *Bacillus pumilus* group (as reported here and by others (Espariz et al., 2016; X. Fu et al., 2021; Liu et al., 2013)), it is likely that strain C1 truly belongs to a *B. safensis* species; however, the complete genome of C1 is not available to confirm this hypothesis.

The most variable region in the nitrilase protein family is the C-terminal tail that forms beta-strands that mediate intersubunit interaction in polymeric structures (Benedik & Sewell, 2018; Thuku et al., 2009). Thus, we associated a phylogenetic tree obtained from the full-length sequences of identified CynDs homologs to an alignment of the C-terminal region (residues 296 to 330) (Fig. 64B). Residues F314, D318, H323 in *B. safensis* CynD are L314, A318, and N323 in the *B. pumilus* protein. Other residues can vary in one of the species but are strictly conserved in the other, for instance, residues Q309 and I325 in *B. safensis* are T309 or N309 and M325 or L325 in *B. pumilus*. Residue 308 can be P or M in *B. safensis* but is strictly D in *B. pumilus* (Fig. 64B). CynD$_{C1}$ has the aminoacids strictly conserved in *B. safensis* supporting the conclusion that C1 belongs to *B. safensis* species. Furthermore, residue 27, outside the C-terminal, is E in *B. safensis* and strain C1 but Q in *B. pumilus*.

Table 18. Identical protein groups (IPG) NCBI accession IDs by strain and species.

Identical Protein Group	Strain	Specie
WP_003215705.1	LLTC96	*B. pumilus*
	NCTC10337	
WP_012010494.1	SAFR-032	

WP_180310545.1	UAMX	
WP_181462014.1	B4127	
	LNTW65	
WP_186299671.1	150a	
	179-D 9B5 HS	
	DE0012	
	DE0037	
	DE0045	
	DE0075	
	DE0078	
	DE0094	
	DE0101	
	DE0146	
	DE0186	
	DE0262	
	DE0264	
	DE0283	
	DE0305	
	DE0317	
	DE0333	
	DE0342	
	DE0470	
	DE0471	
	DE0548	
	DE0560	
	SF-4	
WP_186306833.1	DE0461	
	SH-B9	
WP_186314400.1	DE0035	
	DE0079	
	DE0107	
	DE0154	
	DE0192	
	DE0278	
WP_186325024.1	DE0072	
	DE0119	
WP_189282688.1	EZ-C07	
	LDZX38	
	NMSW10	
WP_189318718.1	LNXM70	
WP_211064195.1	179-K 3C2 HS	
PER-URP-08	PER-URP-08	*B. safensis*
WP_029706059.1	DE0105	
	KCTC 12796BP	
WP_169510666.1	I67	
WP_170825868.1	B4134	

	BRM1	
	FO-36b	
WP_180272414.1	B4129	
	GBSW22	
WP_181566846.1	JG-B5T	
WP_183002030.1	47a_TX	
WP_186318645.1	DE0299	
	F6	
WP_186437300.1	3300	
WP_187470524.1	PgKB20	
WP_196770530.1	U17-1	
	U41	
WP_197172681.1	sami	
WP_214755530.1	ISL-93	

2.3.4.5.- CynD from *B. safensis* PER-URP-08 it is still active at pH 9.

We then went on to characterize some biochemical properties of $\text{CynD}_{\text{PER-URP-08}}$. First, we cloned and expressed recombinant $\text{CynD}_{\text{PER-URP-08}}$ with a C-terminal 6x-His-tag in *E. coli* and determined the basic kinetic constants of the purified recombinant enzyme. Although CynDs are known to be able to adopt different oligomeric states, no evidence of cooperativity was observed in our enzymatic assays (Fig. 65A). Instead, a simple Michaelis-Menten model fit the experimental data adequately. K_m and k_{cat} estimated using this model were 1.93 mmol/L and 6.85 s^{-1} (Fig. 65A, 66).

Due to the volatility of hydrogen cyanide in its protonated HCN state and its pKa of 9.2 (Brüger et al., 2018), bioremediation processes should preferably be carried out at or above pH 9. To test if $\text{CynD}_{\text{PER-URP-08}}$ is active at pHs greater than 8, we tested its activity at pH 9, 10, and 11. Figure 65B shows that recombinant $\text{CynD}_{\text{PER-URP-08}}$ carrying a C-terminal 6x-His tag is active up to pH 9 and inactive at pH 10 and 11. Other wild-type

CynDs have been shown to be active only up to pH 8 (Crum et al., 2016; Jandhyala et al., 2005) and CynD$_{C1}$ with C-terminal 6x-His tag had its activity compromised at pH 9 (Vargas-Serna et al., 2020). The CynD$_{C1}$ and CynD$_{PER-URP-08}$ sequences only differ at five positions: are I18V, S25T, E155D, H305Q and N307Y (first letter correspond to CynD$_{C1}$) with the last two substitutions H305Q and N307Y near the C-terminus.

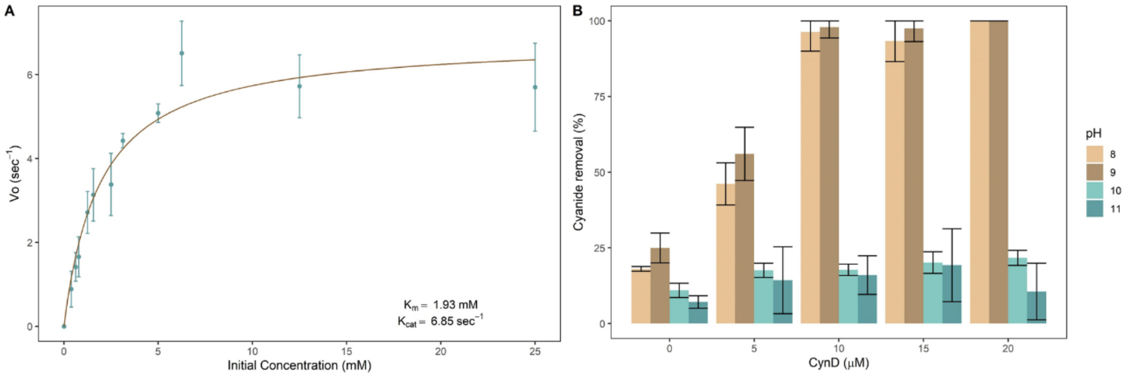


Figure 65. CynD$_{PER-URP-08}$ have similar kinetic constants to other CynD homologues and is still active up to pH 9. A) Plot of CynD$_{PER-URP-08}$ Initial velocity (Vo) versus initial concentration of cyanide adjusted to the Michaelis Menten equation. Km and Kcat constants calculated assuming this model are shown in the graphic. Reactions were done using 500 nM CynD$_{PER-URP-08}$ at pH 8.0, at 30 ºC. B) Percentage of cyanide removal in different pHs using different CynD$_{PER-URP-08}$ concentrations. CynD$_{PER-URP-08}$ showed considerable activity in pH 8 and 9 but not in 10 and 11.

Other studies were able to generate active versions of CynD active at pH 9 by introducing mutations in some conserved positions (K93R; Q86R, E96G, D254G) or by replacing the C-terminal from CynD$_{C1}$ with the C-terminal from CynD from *Pseudomonas stutzeri* (Crum et al., 2015; Wang et al., 2012) (note that wild-type CynD from *P. stutzeri* has not been tested at pH 9).

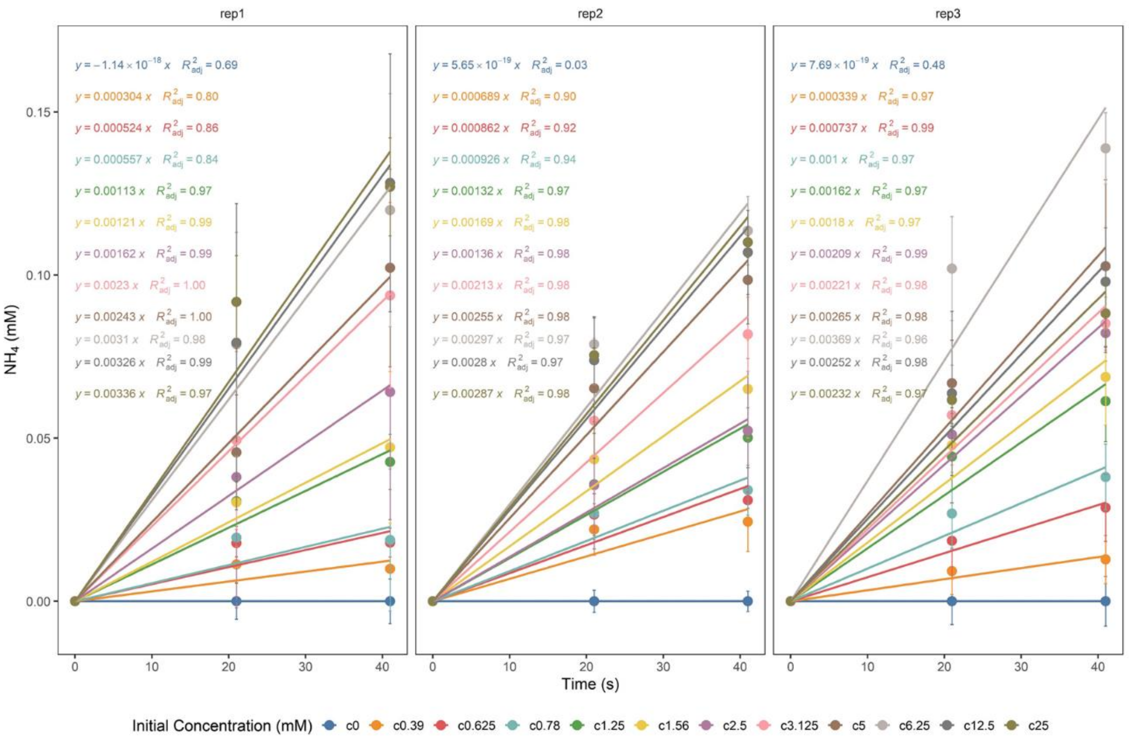


Figure 66. CynD$_{PER-URP-08}$ production of NH$_4$ by time. Linear adjust of the product formation (NH$_4$) by CynD in the first 40 seconds of reaction using different initial concentrations of cyanide.

2.3.4.6.- Alkaline pH reduces the degree of oligomerization of CynD$_{PER-URP-08}$

The oligomerization state of nitrilases have been associated with enzyme activity and stability (Crum et al., 2015; Crum et al., 2015; Crum et al., 2016; Martínková et al., 2015; Park et al., 2016; Wang et al., 2012). In the case of CynDs of CynD$_{C1}$ and CynD from P. stutzeri, mutations in the C-terminal region decrease oligomerization (M. Crum et al., 2016; M. A. N. Crum et al., 2015; Wang et al., 2012). The C-terminal of nitrilases stabilizes the spiral structure through crisscrossed beta sheets in the center of the oligomer (Mulelu et al., 2019; Thuku et al., 2009). Also, pH has been shown to promote higher order oligomerization states of CynDs (D. Jandhyala et al., 2003; Wang et al., 2012); however, the effects in CynD oligomerization at pH greater than 9 have not been reported.

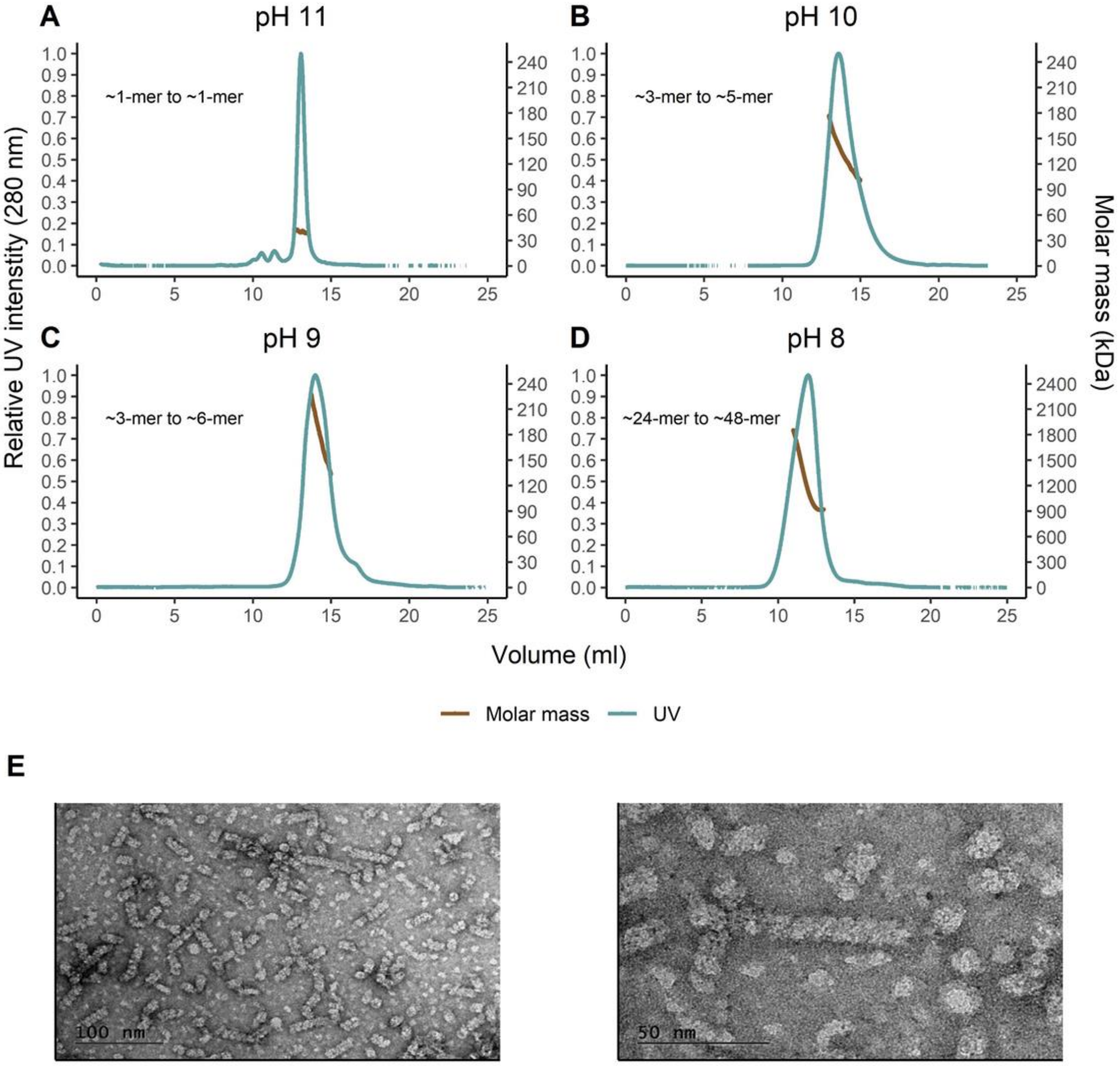


Figure 67. SEC-MALS of CynD$_{PER-URP-08}$ showed that higher pHs reduced its oligomerization states and TEM showed that CynD$_{PER-URP-08}$ presents a helical structure. A-D) Plot of UV intensity/molar mass for CynD$_{PER-URP-08}$ in different pHs. A pattern of decrease the oligomeric state while increasing the pH was observed. E) TEM micrographs at pH 8 in two different magnifications (right and left) showed helical structures of CynD$_{PER-URP-08}$.

Since, CynD$_{\text{PER-URP-08}}$ has differences in C-terminal with respect to other CynDs we used SEC-MALS to compare the oligomerization states of CynD$_{\text{PER-URP-08}}$ at different pHs. As expected, pHs higher than 8 results in smaller sized oligomers. At pH 11 the monomer (38.5 kDa) is the predominant species (Fig. 67A), whereas pH 10 and 9 presented oligomeric states ranging from ~3-mer to ~5-mer (pH 10, 100.85 to 176.34 kDa) and ~4-mer to ~6-mer (pH 9, 133.19 to 226.99 kDa) (Fig. 67B-C). Furthermore, CynD$_{\text{PER-URP-08}}$ presented oligomers ranging from ~24-mer to ~48-mer (918.31 to 1851.39 kDa) at pH 8 (Fig. 67D) in contrast to what was reported for CynD$_{\text{C1}}$ at pH 8 which forms an 18-mer spiral (D. Jandhyala et al., 2003). These differences could be a result of the differences in aminoacid sequence between CynD$_{\text{PER-URP-08}}$ and CynD$_{\text{C1}}$ or due to the presence of the C-terminal 6x-His tag in CynD$_{\text{PER-URP-08}}$. Experiments with CynD$_{\text{C1}}$ were carried out with untagged protein or with protein carrying an N-terminal 6x-His tag (Crum et al., 2015; Jandhyala et al., 2003; Park et al., 2016; Wang et al., 2012). Electron micrographs of negatively stained CynD$_{\text{PER-URP-08}}$ at pH 8 showed spirals of different sizes supporting the conclusion that CynD$_{\text{PER-URP-08}}$ at this pH adopts a range of different oligomerization states (Fig. 67E).

2.3.4.7.- Expression of CynD$_{\text{PER-URP-08}}$ from *B. safensis* PER-URP-08 is induced in the presence of cyanide

Some previous studies have considered the possibility that CynD gene expression is regulated by cyanide, but this point remains unclear (D. Jandhyala et al., 2003). To address this question, we exposed *B. safensis* PER-URP-08 to 100 ppm CN$^-$ (in the form of 38.5 mmol/L NaCN) at 30 °C for 4 h without agitation and the mRNA levels of cynD were measured and compared with the levels observed in cells grown in the absence of CN-.

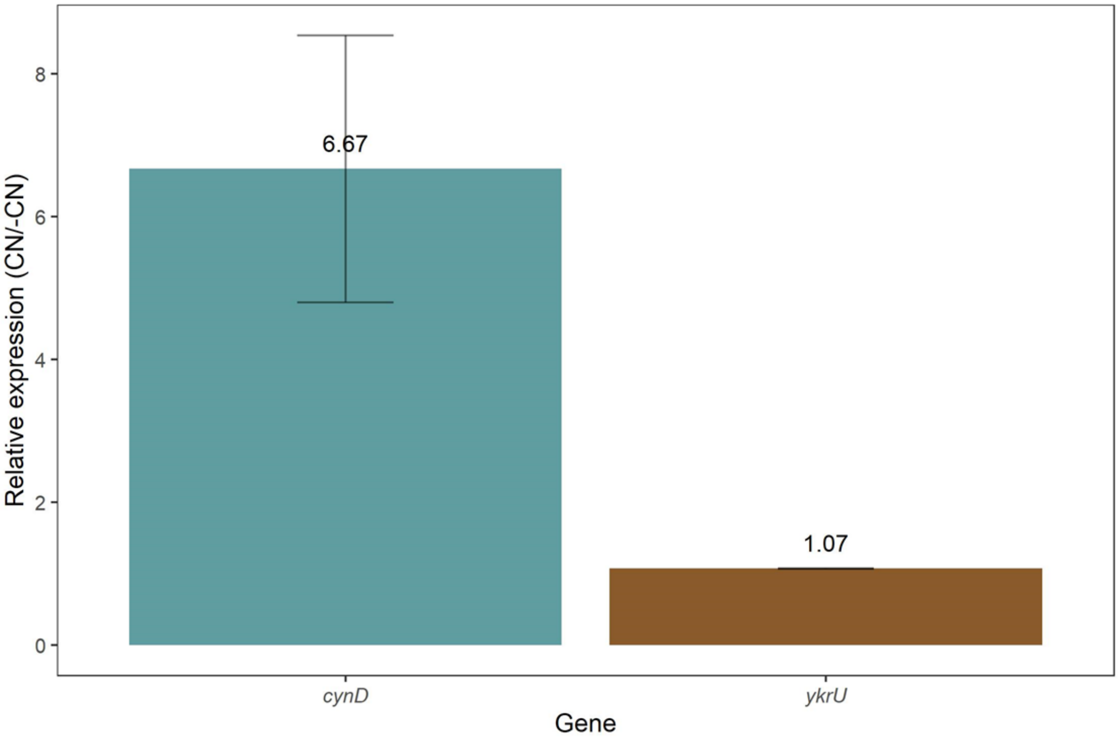


Figure 68. *cynD*$_{\text{PER-URP-08}}$ but not *ykrU* is induced in the presence of cyanide. Relative expression measured by RT-qPCR showed that when *Bacillus safensis* is in presence of cyanide the RNA levels of *cynD* are 6.67-fold greater than when in absence of cyanide. In contrast, other nitrilase gene (*ykrU*) have the same RNA levels in presence or absence of cyanide.

We observed a 6.7-fold increase in expression of cynD in the presence of cyanide (Fig. 57). To evaluate if this overexpression is specific for cynD nitrilase and not to other nitrilases of *B. safensis* PER-URP-08, we also measured the mRNA levels of *ykrU* that also possesses a CN_hydrolase domain. We did not observe differences in *ykrU* expression in the presence and absence of cyanide. To our knowledge, this is the first report showing induction in the expression of *cynD* in the presence of cyanide. This could possibly be a physiological response of the bacteria in order to protect itself from the toxic effects of the compound, but further studies are necessary to understand the molecular mechanisms more fully behind this response.

2.3.5.- CONCLUSIONS

Here we report the isolation and the genome sequences of three cyanide-degrading *Bacillus* strains obtained from water in contact with mine tailings in Lima – Peru. They were phylogenetically classified and named *Bacillus licheniformis* PER-URP-12, Bacillus subtilis PER-URP-17 and *Bacillus safensis* PER-URP-08. Comparative genomic analyses indicate that some strains currently classified as *B. pumilus* with publicly available genomes should be reclassified as *Bacillus altitudinis* (strains TUAT1, MTCB 6033, SH-B11, and C4). Furthermore, we propose that strains CH144a_4T and 145 should be classified belonging a new species distinct from *B. pumilus*, *B. safensis*, or *B. altitudinis*.

We propose that in *B. licheniformis* PER-URP-12 and *B. subtilis* PER-URP-17 rhodaneses (table 17) are possibly the enzymes that confer cyanide degradation capacities to these strains. In the case of *B. safensis* PER-URP-08, we suggest that EGI07_08135 codes for an ortholog of cyanide dihydratase, CynD, that imparts the cyanide-degradation ability to this strain.

We found that while no *B. altitudinis* strains code for CynD orthologs, some *B. pumilus* and *B. safensis* strains present CynD orthologous sequences. CynD from *B. pumilus* and *B. safensis* have high identity (> 97%); however conserved differences in the C-terminus allow us to differentiate between CynD from *B. safensis* or *B. pumilus* (at least in the analyzed genomes). Additionally, sequence analysis of the previously described CynD from strain C1 (CynD_{C1}), named *B. pumilus* CynD in the literature, is more closely related to CynDs from *B. safensis* than from *B. pumilus*. We characterized some aspects of CynD from *B. safensis* PER-URP-08 ($\text{CynD}_{\text{PER-URP-08}}$) corroborating what was described for CynDs from other species and adding new knowledge about these enzymes. First, enzymatic assays with $\text{CynD}_{\text{PER-URP-08}}$ found no evidence of cooperativity despite the known oligomerization patterns of these enzymes. Second, K_m and k_{cat} of $\text{CynD}_{\text{PER-URP-08}}$ were 1.93 mmol/L and 6.65 s-1, respectively. Third, despite the fact that $\text{CynD}_{\text{PER-URP-08}}$ and CynD_{C1} only differ in five positions, $\text{CynD}_{\text{PER-URP-08}}$ retains almost the same activity at pH 9 that it exhibits at pH 8 whereas CynD_{C1} has been reported to be almost inactive at pH 9. Fourth, as pH is known to influence the oligomerization of CynDs, we reported that at pH 8, $\text{CynD}_{\text{PER-URP-08}}$ forms spirals made up of an estimated ~24 to ~48 subunits showing that several oligomeric states are present in this pH. This is different compared with CynD_{C1} that was reported to mainly form oligomers of 18 subunits at this pH. Moreover, at pH 11, the $\text{CynD}_{\text{PER-URP-08}}$ monomer was observed. Finally, we showed for the first time that the abundance of $\text{CynD}_{\text{PER-URP-08}}$ transcripts increases 6-fold when bacterial cultures are exposed to CN^-.

Altogether, the results we reported here warrant further investigation to explore the potential application of *B. safensis* PER-URP-08 and CynD$_{\text{PER-URP-08}}$ in cyanide bioremediation processes.

2.3.6.- DATA AVAILABILITY

The final genomes assemblies are available in IMG/M (Chen et al., 2021) and GenBank/NCBI (Benson et al., 2013) databases under the accessions numbers: 2818991268, 2818991267, 2818991266 and RSEW00000000.1, RSEY00000000.1, RSEX00000000.1, respectively for *Bacillus safensis* PER-URP-08, *Bacillus licheniformis* PER-URP-12, *Bacillus subtilis* PER-URP-17.

2.3.7.- REFERENCES

Acosta, M., Beard, S., Ponce, J., Vera, M., Mobarec, J. C., & Jerez, C. A. (2005). Identification of putative sulfurtransferase genes in the extremophilic Acidithiobacillus ferrooxidans ATCC 23270 genome: Structural and functional characterization of the proteins. OMICS A Journal of Integrative Biology, 9(1), 13–29. https://doi.org/10.1089/omi.2005.9.13

Adams, H., Teertstra, W., Koster, M., & Tommassen, J. (2002). PspE (phage-shock protein E) of Escherichia coli is a rhodanese. FEBS Letters, 518(1–3), 173–176. https://doi.org/10.1016/S0014-5793(02)02695-9

Akinpelu, E. A., Adetunji, A. T., Ntwampe, S. K. O., Nchu, F., & Mekuto, L. (2018). Performance of fusarium oxysporum EKT01/02 isolate in cyanide biodegradation system. Environmental Engineering Research, 23(2), 223–227. https://doi.org/10.4491/eer.2017.154

Al-Badri, B. A. S., Al-Maawali, S. S., Al-Balushi, Z. M., Al-Mahmooli, I. H., Al-Sadi, A. M., & Velazhahan, R. (2020). Cyanide degradation and antagonistic potential of endophytic Bacillus subtilis strain BEB1 from Bougainvillea spectabilis Willd. All Life, 13(1), 92–98. https://doi.org/10.1080/26895293.2020.1728393

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Atkinson, A. (1975). Bacterial Cyanide Detoxijkation. Biotechnology and Bioengineering, 17, 457–460.

Benedik, M. J., & Sewell, B. T. (2018). Cyanide-degrading nitrilases in nature. Journal of General and Applied Microbiology, 64(2), 90–93. https://doi.org/10.2323/jgam.2017.06.002

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic Acids Research, 41(D1), 36–42.

https://doi.org/10.1093/nar/gks1195

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M. F., Lió, P., Crescenzi, P., Fani, R., & Fondi, M. (2015). MeDuSa: A multi-draft based scaffolder. Bioinformatics, 31(15), 2443–2451. https://doi.org/10.1093/bioinformatics/btv171

Brüger, A., Fafilek, G., Restrepo B., O. J., & Rojas-Mendoza, L. (2018). On the volatilisation and decomposition of cyanide contaminations from gold mining. Science of the Total Environment, 627, 1167–1173. https://doi.org/10.1016/j.scitotenv.2018.01.320

Castric, P. A., & Strobel, G. A. (1969). Cyanide metabolism by Bacillus megaterium. Journal of Biological Chemistry, 244(15), 4089–4094. https://doi.org/10.1016/s0021-9258(17)36388-3

Chen, I. M. A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., Roux, S., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2021). The IMG/M data management and analysis system v.6.0: New tools and advanced capabilities. Nucleic Acids Research, 49(D1), D751–D763. https://doi.org/10.1093/nar/gkaa939

Cipollone, R., Ascenzi, P., Frangipani, E., & Visca, P. (2006). Cyanide detoxification by recombinant bacterial rhodanese. Chemosphere, 63(6), 942–949.

https://doi.org/10.1016/j.chemosphere.2005.09.048

Cipollone, R., Frangipani, E., Tiburzi, F., Imperi, F., Ascenzi, P., & Visca, P. (2007). Involvement of Pseudomonas aeruginosa rhodanese in protection from cyanide toxicity. Applied and Environmental Microbiology, 73(2), 390–398. https://doi.org/10.1128/AEM.02143-06

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Coil, D., Jospin, G., & Darling, A. E. (2015). A5-miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics, 31(4), 587–589. https://doi.org/10.1093/bioinformatics/btu661

Crum, M. A., Park, J. M., Mulelu, A. E., Sewell, B. T., & Benedik, M. J. (2015). Probing C-terminal interactions of the Pseudomonas stutzeri cyanide-degrading CynD protein. Applied Microbiology and Biotechnology, 99(7), 3093–3102. https://doi.org/10.1007/s00253-014-6335-x

Crum, M. A., Park, J. M., Sewell, B. T., & Benedik, M. J. (2015). C-terminal hybrid mutant of Bacillus pumilus cyanide dihydratase dramatically enhances thermal stability and pH tolerance by reinforcing oligomerization. Journal of Applied Microbiology, 118(4), 881–889. https://doi.org/10.1111/jam.12754

Crum, M. A., Trevor, B., & Benedik, M. (2016). Bacillus pumilus cyanide dihydratase mutants with higher catalytic activity. Frontiers in Microbiology, 7(AUG), 1–10. https://doi.org/10.3389/fmicb.2016.01264

Dash, R. R., Gaur, A., & Balomajumder, C. (2009). Cyanide in industrial wastewaters and its removal: A review on biotreatment. Journal of Hazardous Materials, 163(1), 1–11. https://doi.org/10.1016/j.jhazmat.2008.06.051

Dumestre, A., Chone, T., Portal, J. M., Gerard, M., & Berthelin, J. (1997). Cyanide degradation under alkaline conditions by a strain of Fusarium solani isolated from contaminated soils. Applied and Environmental Microbiology, 63(7), 2729–2734. https://doi.org/10.1128/aem.63.7.2729-2734.1997

Ellens, K. W., Richardson, L. G. L., Frelin, O., Collins, J., Ribeiro, C. L., Hsieh, Y. F., Mullen, R. T., & Hanson, A. D. (2015). Evidence that glutamine transaminase and omega-amidase potentially act in tandem to close the methionine salvage cycle in bacteria and plants. Phytochemistry, 113, 160–169. https://doi.org/10.1016/j.phytochem.2014.04.012

Espariz, M., Zuljan, F. A., Esteban, L., & Magni, C. (2016). Taxonomic identity resolution of highly phylogenetically related strains and selection of phylogenetic markers by using genome-scale methods: The bacillus pumilus group case. PLoS ONE, 11(9), 1–17. https://doi.org/10.1371/journal.pone.0163098

Fry, W. E., & Millar, R. L. (1972). Cyanide degradion by an enzyme from Stemphylium loti. Archives of Biochemistry and Biophysics, 151(2), 468–474. https://doi.org/10.1016/0003-9861(72)90523-1

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Fu, X., Gong, L., Liu, Y., Lai, Q., Li, G., & Shao, Z. (2021). Bacillus pumilus Group Comparative Genomics: Toward Pangenome Features, Diversity, and Marine Environmental Adaptation. Frontiers in Microbiology, 12(May), 1–16. https://doi.org/10.3389/fmicb.2021.571212

Gioia, J., Yerrapragada, S., Qin, X., Jiang, H., Igboeli, O. C., Muzny, D., Dugan-Rocha, S., Ding, Y., Hawes, A., Liu, W., Perez, L., Kovar, C., Dinh, H., Lee, S., Nazareth, L., Blyth, P., Holder, M., Buhay, C., Tirumalai, M. R., … Weinstock, G. M. (2007). Paradoxical DNA repair and peroxide resistance gene conservation in Bacillus pumilus SAFR-032. PLoS ONE, 2(9). https://doi.org/10.1371/journal.pone.0000928

Heidelberg, J. F., Elsen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., Gill, S. R., Nelson, K. E., Read, T. D., Tettelin, H., Richardson, D., Ermolaeva, M. D., Vamathevan, J., Bass, S., Halving, Q., … Fraser, C. M. (2000). DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature, 406(6795), 477–483. https://doi.org/10.1038/35020000

Hendry-Hofer, T. B., Ng, P. C., Witeof, A. E., Mahon, S. B., Brenner, M., Boss, G. R., & Bebarta, V. S. (2019). A Review on Ingested Cyanide: Risks, Clinical Presentation, Diagnostics, and Treatment Challenges. Journal of Medical Toxicology, 15(2), 128–133. https://doi.org/10.1007/s13181-018-0688-y

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution, 35(2), 518–522. https://doi.org/10.1093/molbev/msx281

Ibrahim, K. K., Syed, M. A., Shukor, M. Y., & Ahmad, S. A. (2015). Biological remediation of cyanide: A review. Biotropia, 22(2), 151–163. https://doi.org/10.11598/btb.2015.22.2.393

Idris, E. S. E., Iglesias, D. J., Talon, M., & Borriss, R. (2007). Tryptophan-dependent production of Indole-3-Acetic Acid (IAA) affects level of plant growth promotion by Bacillus amyloliquefaciens FZB42. Molecular Plant-Microbe Interactions, 20(6), 619–626. https://doi.org/10.1094/MPMI-20-6-0619

Ingvorsen, K., Hojer-Pedersen, B., & Godtfredsen, S. E. (1991). Novel cyanide-hydrolyzing enzyme from Alcaligenes xylosoxidans subsp. denitrificans. Applied and Environmental Microbiology, 57(6), 1783–1789. https://doi.org/10.1128/aem.57.6.1783-1789.1991

Itakorode, B., Okonji, R., Adedeji, O., Torimiro, O., Famakinwa, T., & Chukwuejim, C. (2019). Isolation, screening and optimization of Bacillus cereus for a thiosuphate sulphur transferase production. Journal of Chemical and Pharmaceutical Sciences, 12(03), 79–84. https://doi.org/10.30558/jchps.20191203003

Jandhyala, D., Berman, M., Meyers, P. R., Sewell, B. T., Willson, R. C., & Benedik, M. J. (2003). CynD, the cyanide dihydratase from Bacillus pumilus: Gene cloning and structural studies. Applied and Environmental Microbiology, 69(8), 4794–4805. https://doi.org/10.1128/AEM.69.8.4794-4805.2003

Jandhyala, D. M., Willson, R. C., Sewell, B. T., & Benedik, M. J. (2005). Comparison of cyanide-degrading nitrilases. Applied Microbiology and Biotechnology, 68(3), 327–335. https://doi.org/10.1007/s00253-005-1903-8

Javaheri Safa, Z., Aminzadeh, S., Zamani, M., & Motallebi, M. (2017). Significant increase in cyanide degradation by Bacillus sp. M01 PTCC 1908 with response surface methodology optimization. AMB Express, 7(1). https://doi.org/10.1186/s13568-017-0502-2

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kosiol, C., & Goldman, N. (2005). Different versions of the dayhoff rate matrix. Molecular Biology and Evolution, 22(2), 193–199. https://doi.org/10.1093/molbev/msi005

Kuyucak, N., & Akcil, A. (2013). Cyanide and removal options from effluents in gold mining and metallurgical processes. Minerals Engineering, 50–51, 13–29. https://doi.org/10.1016/j.mineng.2013.05.027

Leavesley, H. B., Li, L., Prabhakaran, K., Borowitz, J. L., & Isom, G. E. (2008). Interaction of cyanide and nitric oxide with cytochrome c oxidase: Implications for acute cyanide toxicity. Toxicological Sciences, 101(1), 101–111. https://doi.org/10.1093/toxsci/kfm254

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liu, Y., Lai, Q., Dong, C., Sun, F., Wang, L., Li, G., & Shao, Z. (2013). Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis. PLoS ONE, 8(11), 1–11. https://doi.org/10.1371/journal.pone.0080097

Martínková, L., Veselá, A. B., Rinágelová, A., & Chmátal, M. (2015). Cyanide hydratases and cyanide dihydratases: emerging tools in the biodegradation and biodetection of cyanide. Applied Microbiology and Biotechnology, 99(21), 8875–8882. https://doi.org/10.1007/s00253-015-6899-0

Mekuto, L., Jackson, V. A., & Obed Ntwampe, S. K. (2014). Biodegradation of Free Cyanide Using Bacillus Sp. Consortium Dominated by Bacillus Safensis, Lichenformis and Tequilensis Strains: A Bioprocess Supported Solely with Whey. Journal of Bioremediation & Biodegradation, 05(02). https://doi.org/10.4172/2155-6199.s18-004

Melino, S., Cicero, D. O., Orsale, M., Forlani, F., Pagani, S., & Paci, M. (2003). Azotobacter vinelandii rhodanese Selenium loading and ion interaction studies. European Journal of Biochemistry, 270(20), 4208–4215. https://doi.org/10.1046/j.1432-1033.2003.03818.x

Meyers, P. R., Rawlings, D. E., Woods, D. R., & Lindsey, G. G. (1993). Isolation and characterization of a cyanide dihydratase from Bacillus pumilus C1. Journal of

Bacteriology, 175(19), 6105–6112. https://doi.org/10.1128/jb.175.19.6105-6112.1993

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mudder, T. I., Botz, M. M., & Akçil, A. (2004). Cyanide and society: A critical review. The European Journal of Mineral Processing and Environmental Protection, 4(1), 62–74.

Mulelu, A. E., Kirykowicz, A. M., & Woodward, J. D. (2019). Cryo-EM and directed evolution reveal how Arabidopsis nitrilase specificity is influenced by its quaternary structure. Communications Biology, 2(1), 1–11. https://doi.org/10.1038/s42003-019-0505-4

Omura, H., Ikemoto, M., Kobayashi, M., Shimizu, S., Yoshida, T., & Nagasawa, T. (2003). Purification, characterization and gene cloning of thermostable O-acetyl-L-homoserine sulfhydrylase forming γ-cyano-α-aminobutyric acid. Journal of Bioscience and Bioengineering, 96(1), 53–58. https://doi.org/10.1016/S1389-1723(03)90096-X

Pace, H. C., & Brenner, C. (2001). The nitrilase superfamily: Classification, structure and function. Genome Biology, 2(1), 1–9. https://doi.org/10.1186/gb-2001-2-1-reviews0001

Park, J. M., Ponder, C. M., Sewell, B. T., & Benedik, M. J. (2016). Residue Y70 of the nitrilase cyanide dihydratase from Bacillus pumilus is critical for formation and activity of the spiral oligomer. Journal of Microbiology and Biotechnology, 26(12), 2179–2183. https://doi.org/10.4014/jmb.1606.06035

Ping Wang; Hans D. VanEtten. (1992). Cloning and properties of a cyanide hydratase gene from the phytopathogenic fungus Gloeocercospora sorghi. Biochemical and Biophysical Research Communications, 187(2), 1048–1054. https://www.sciencedirect.com/science/article/abs/pii/0006291X92913038

Ravot, G., Casalot, L., Ollivier, B., Loison, G., & Magot, M. (2005). RdlA, a new gene encoding a rhodanese-like protein in Halanaerobium congolense and other thiosulfate-reducing anaerobes. Research in Microbiology, 156(10), 1031–1038. https://doi.org/10.1016/j.resmic.2005.05.009

Rinágelová, A., Kaplan, O., Veselá, A. B., Chmátal, M., Křenková, A., Plíhal, O., Pasquarelli, F., Cantarella, M., & Martínková, L. (2014). Cyanide hydratase from Aspergillus niger K10: Overproduction in Escherichia coli, purification, characterization and use in continuous cyanide degradation. Process Biochemistry, 49(3), 445–450.

https://doi.org/10.1016/j.procbio.2013.12.008

Satomi, M., La Duc, M. T., & Venkateswaran, K. (2006). Bacillus safensis sp.nov., isolated from spacecraft and assembly-facility surfaces. International Journal of Systematic and Evolutionary Microbiology, 56(8), 1735–1740. https://doi.org/10.1099/ijs.0.64189-0

Sekowska, A., & Danchin, A. (2002). The methionine salvage pathway in Bacillus subtilis. BMC Microbiology, 2, 1–14. https://doi.org/10.1186/1471-2180-2-8

Sexton, A. C., & Howlett, B. J. (2000). Characterisation of a cyanide hydratase gene in the phytopathogenic fungus Leptosphaeria maculans. *Molecular and General Genetics*, *263*(3), 463–470. https://doi.org/10.1007/s004380051190

Stepanov, V. G., Tirumalai, M. R., Montazari, S., Checinska, A., Venkateswaran, K., & Fox, G. E. (2016). Bacillus pumilus SAFR-032 genome revisited: Sequence update and re-annotation. *PLoS ONE*, *11*(6), 1–11. https://doi.org/10.1371/journal.pone.0157331

Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569

Thuku, R. N., Brady, D., Benedik, M. J., & Sewell, B. T. (2009). Microbial nitrilases: Versatile, spiral forming, industrial enzymes. *Journal of Applied Microbiology*, *106*(3), 703–727. https://doi.org/10.1111/j.1365-2672.2008.03941.x

Vargas-Serna, C. L., Carmona-Orozco, M. L., & Panay, A. J. (2020). Biodegradation of cyanide using recombinant Escherichia coli expressing Bacillus pumilus cyanide dihydratase. *Revista Colombiana de Biotecnología*, *22*(1), 27–35. https://doi.org/10.15446/rev.colomb.biote.v22n1.79559

Wang, L., Watermeyer, J. M., Mulelu, A. E., Sewell, B. T., & Benedik, M. J. (2012). Engineering pH-tolerant mutants of a cyanide dihydratase. *Applied Microbiology and Biotechnology*, *94*(1), 131–140. https://doi.org/10.1007/s00253-011-3620-9

Watanabe, A., Yano, K., Ikebukuro, K., & Karube, I. (1998). Cloning and expression of a gene encoding cyanidase from Pseudomonas stutzeri AK61. *Applied Microbiology and Biotechnology*, *50*(1), 93–97. https://doi.org/10.1007/s002530051261

Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., Nusbaum, C., Lander, E. S., Maccallum, I., & Jaffe, D. B. (2014). Comprehensive variation discovery in single human genomes. *Nature Genetics*, *46*(12), 1350–1355. https://doi.org/10.1038/ng.3121

Williams, H. J., & Edwards, T. G. (1980). Estimation of cyanide with alkaline picrate.

Journal of the Science of Food and Agriculture, *31*(1), 15–22. https://doi.org/10.1002/jsfa.2740310104

2.3.8.- ACKNOWLEDGEMENTS

We are very grateful to the Ricardo Palma University High-Performance Computational Cluster (URPHPC) managers Gustavo Adolfo Abarca Valdiviezo and Roxana Paola Mier Hermoza at the Ricardo Palma Informatic Department (OFICIC) for their contribution in programs and remote use configuration of URPHPC. We also thank Germán Sgro for help with negative staining.

2.3.9.- FUNDING INFORMATION

Funding for this work was provided by São Paulo Research Foundation (FAPESP) student fellowship 2015/13318-4 to S.J.A and research grant 2017/17303-7 to C.S.F. The work was also supported by Coordination for the Improvement of Higher Education Personnel (CAPES) research grant 3385/2013 to A.M.D.S. P.M.P. received fellowships from CAPES (88882.160114/2017-1). C.S.F., and J.C.S., A.M.D.S. received research fellowship awards from the National Council for Scientific and Technological Development (CNPq). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.3.10.- AUTHORS CONTRIBUTIONS

Conceptualization: S.J.A., A.G.S., A.M.D.S., Methodology: S.J.A., D.Z.S., A.C.P., M.B.R., C.M.P., L.F.M., P.M.P., C.M. Computing resources: M.Q.A., J.C.S., Data curation: S.J.A., D.Z.S., A.C.P., C.M.P. Formal analysis: S.J.A., D.Z.S., A.C.P., C.M.P., C.S.F. Visualization: S.J.A., D.Z.S. Writing – original draft preparation: S.J.A., Writing – review and editing: S.J.A., J.C.S., C.S.F., A.M.D.S. Supervision: S.J.A., A.G.S., M.Q.A., C.S.F., A.M.D.S. Funding acquisition: A.G.S., M.Q.A., C.S.F., J.C.S., A.M.D.S. All authors read, provided critical review, and approved the final manuscript.

2.3.11.- CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

2.4.- REFERENCES

Akinpelu, E. A., Adetunji, A. T., Ntwampe, S. K. O., Nchu, F., & Mekuto, L. (2018). Performance of fusarium oxysporum EKT01/02 isolate in cyanide biodegradation system. *Environmental Engineering Research*, *23*(2), 223–227. https://doi.org/10.4491/eer.2017.154

Angove, J. E., & Acar, S. (2016). Metallurgical Test Work. In *Gold Ore Processing*. Elsevier B.V. https://doi.org/10.1016/b978-0-444-63658-4.00008-6

Black, G. W., Brown, N. L., Perry, J. J. B., Randall, P. D., Turnbull, G., & Zhang, M. (2015). A high-throughput screening method for determining the substrate scope of nitrilases. *Chemical Communications*, *51*(13), 2660–2662. https://doi.org/10.1039/c4cc06021k

Brenner, C. (2002). Catalysis in the nitrilase superfamily. *Current Opinion in Structural Biology*, *12*(6), 775–782. https://doi.org/10.1016/S0959-440X(02)00387-1

Chhiba-Govindjee, V. P., van der Westhuyzen, C. W., Bode, M. L., & Brady, D. (2019). Bacterial nitrilases and their regulation. *Applied Microbiology and Biotechnology*, *103*(12), 4679–4692. https://doi.org/10.1007/s00253-019-09776-1

Cooper, C. E., & Brown, G. C. (2008). The inhibition of mitochondrial cytochrome oxidase by the gases carbon monoxide, nitric oxide, hydrogen cyanide and hydrogen sulfide: Chemical mechanism and physiological significance. *Journal of Bioenergetics and Biomembranes*, *40*(5), 533–539. https://doi.org/10.1007/s10863-008-9166-6

Crane, F. L., Sun, I. L., Barr, R., & Löw, H. (1991). Electron and proton transport across the plasma membrane. In *Journal of Bioenergetics and Biomembranes* (Vol. 23, Issue 5). https://doi.org/10.1007/BF00786001

Dash, R. R., Gaur, A., & Balomajumder, C. (2009). Cyanide in industrial wastewaters and its removal: A review on biotreatment. *Journal of Hazardous Materials*, *163*(1), 1–11. https://doi.org/10.1016/j.jhazmat.2008.06.051

Dumestre, A., Chone, T., Portal, J. M., Gerard, M., & Berthelin, J. (1997). Cyanide degradation under alkaline conditions by a strain of Fusarium solani isolated from contaminated soils. *Applied and Environmental Microbiology*, *63*(7), 2729–2734. https://doi.org/10.1128/aem.63.7.2729-2734.1997

Eisler, R. (1991). Cyanide hazards to fish, wildlife, and invertebrates: a synoptic review-U.S. Fish Wildl. Serv. *Biological Report*, *85*(1.23), 1–65. http://catalogue.nla.gov.au/Record/4005011

Fernandes, B. C. M., Mateo, C., Kiziak, C., Chmura, A., Wacker, J., Van Rantwijk, F., Stolz, A., & Sheldon, R. A. (2006). Nitrile hydratase activity of a recombinant nitrilase. *Advanced Synthesis and Catalysis*, *348*(18), 2597–2603. https://doi.org/10.1002/adsc.200600269

Ferus, M., Rimmer, P., Cassone, G., Knízek, A., Civiš, S., Šponer, J. E., Ivanek, O., Šponer, J., Saeidfirozeh, H., Kubelík, P., Dudzák, R., Petera, L., Juha, L., Pastorek, A., Křivková, A., & Krůs, M. (2020). One-Pot Hydrogen Cyanide-Based Prebiotic Synthesis of Canonical Nucleobases and Glycine Initiated by High-Velocity Impacts on Early Earth. *Astrobiology*, *20*(12), 1476–1488. https://doi.org/10.1089/ast.2020.2231

Fry, W. E., & Millar, R. L. (1972). Cyanide degradion by an enzyme from Stemphylium loti. *Archives of Biochemistry and Biophysics*, *151*(2), 468–474. https://doi.org/10.1016/0003-9861(72)90523-1

Gijzen, H. J., Bernal, E., & Ferrer, H. (2000). Cyanide toxicity and cyanide degradation in anaerobic wastewater treatment. *Water Research*, *34*(9), 2447–2454. https://doi.org/10.1016/S0043-1354(99)00418-2

Gong, J. S., Lu, Z. M., Li, H., Shi, J. S., Zhou, Z. M., & Xu, Z. H. (2012). Nitrilases in nitrile biocatalysis: Recent progress and forthcoming research. *Microbial Cell Factories*, *11*, 1–18. https://doi.org/10.1186/1475-2859-11-142

Hendry-Hofer, T. B., Ng, P. C., Witeof, A. E., Mahon, S. B., Brenner, M., Boss, G. R., & Bebarta, V. S. (2019). A Review on Ingested Cyanide: Risks, Clinical Presentation, Diagnostics, and Treatment Challenges. *Journal of Medical Toxicology*, *15*(2), 128–133. https://doi.org/10.1007/s13181-018-0688-y

Ingvorsen, K., Hojer-Pedersen, B., & Godtfredsen, S. E. (1991). Novel cyanide-hydrolyzing enzyme from Alcaligenes xylosoxidans subsp. denitrificans. *Applied and Environmental Microbiology*, *57*(6), 1783–1789. https://doi.org/10.1128/aem.57.6.1783-1789.1991

Jandhyala, D., Berman, M., Meyers, P. R., Sewell, B. T., Willson, R. C., & Benedik, M. J. (2003). CynD, the cyanide dihydratase from Bacillus pumilus: Gene cloning and structural studies. *Applied and Environmental Microbiology*, *69*(8), 4794–4805. https://doi.org/10.1128/AEM.69.8.4794-4805.2003

Johnson, C. A. (2015). The fate of cyanide in leach wastes at gold mines: An environmental perspective. *Applied Geochemistry*, *57*, 194–205. https://doi.org/10.1016/j.apgeochem.2014.05.023

Komeda, H., Hori, Y., Kobayashi, M., & Shimizu, S. (1996). Transcriptional regulation of the Rhodococcus rhodochrous J1 nitA gene encoding a nitrilase. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(20), 10572–

10577. https://doi.org/10.1073/pnas.93.20.10572

Kuyucak, N., & Akcil, A. (2013). Cyanide and removal options from effluents in gold mining and metallurgical processes. *Minerals Engineering*, *50–51*, 13–29. https://doi.org/10.1016/j.mineng.2013.05.027

Leavesley, H. B., Li, L., Prabhakaran, K., Borowitz, J. L., & Isom, G. E. (2008). Interaction of cyanide and nitric oxide with cytochrome c oxidase: Implications for acute cyanide toxicity. *Toxicological Sciences*, *101*(1), 101–111. https://doi.org/10.1093/toxsci/kfm254

Menor Salván, C., Bouza, M., Fialho, D. M., Burcar, B. T., Fernández, F. M., & Hud, N. V. (2020). Prebiotic Origin of Pre-RNA Building Blocks in a Urea “Warm Little Pond” Scenario. *ChemBioChem*, *21*(24), 3504–3510. https://doi.org/10.1002/cbic.202000510

Meyers, P. R., Rawlings, D. E., Woods, D. R., & Lindsey, G. G. (1993). Isolation and characterization of a cyanide dihydratase from Bacillus pumilus C1. *Journal of Bacteriology*, *175*(19), 6105–6112. https://doi.org/10.1128/jb.175.19.6105-6112.1993

Mudder, T. I., Botz, M. M., & Akçil, A. (2004). Cyanide and society: A critical review. *The European Journal of Mineral Processing and Environmental Protection*, *4*(1), 62–74.

Mulelu, A. E., Kirykowicz, A. M., & Woodward, J. D. (2019). Cryo-EM and directed evolution reveal how Arabidopsis nitrilase specificity is influenced by its quaternary structure. *Communications Biology*, *2*(1), 1–11. https://doi.org/10.1038/s42003-

019-0505-4

Pace, H. C., & Brenner, C. (2001). The nitrilase superfamily: Classification, structure and function. *Genome Biology*, *2*(1), 1–9. https://doi.org/10.1186/gb-2001-2-1-reviews0001

Ping Wang; Hans D. VanEtten. (1992). Cloning and properties of a cyanide hydratase gene from the phytopathogenic fungus Gloeocercospora sorghi. *Biochemical and Biophysical Research Communications*, *187*(2), 1048–1054. https://www.sciencedirect.com/science/article/abs/pii/0006291X92913038

Rinágelová, A., Kaplan, O., Veselá, A. B., Chmátal, M., Křenková, A., Plíhal, O., Pasquarelli, F., Cantarella, M., & Martínková, L. (2014). Cyanide hydratase from Aspergillus niger K10: Overproduction in Escherichia coli, purification, characterization and use in continuous cyanide degradation. *Process Biochemistry*, *49*(3), 445–450. https://doi.org/10.1016/j.procbio.2013.12.008

Robertson, D. E., Chaplin, J. A., DeSantis, G., Podar, M., Madden, M., Chi, E., Richardson, T., Milan, A., Miller, M., Weiner, D. P., Wong, K., McQuaid, J., Farwell, B., Preston, L. A., Tan, X., Snead, M. A., Keller, M., Mathur, E., Kretz, P. L., … Short, J. M. (2004). Exploring Nitrilase Sequence Space for Enantioselective Catalysis. *Applied and Environmental Microbiology*, *70*(4), 2429–2436. https://doi.org/10.1128/AEM.70.4.2429-2436.2004

Sewell, B. T., Berman, M. N., Meyers, P. R., Jandhyala, D., & Benedik, M. J. (2003). The cyanide degrading nitrilase from Pseudomonas stutzeri AK61 is a two-fold symmetric, 14-subunit spiral. *Structure*, *11*(11), 1413–1422. https://doi.org/10.1016/j.str.2003.10.005

Sexton, A. C., & Howlett, B. J. (2000). Characterisation of a cyanide hydratase gene in the phytopathogenic fungus Leptosphaeria maculans. *Molecular and General Genetics*, *263*(3), 463–470. https://doi.org/10.1007/s004380051190

Stolz, A., Eppinger, E., Sosedov, O., & Kiziak, C. (2019). Comparative analysis of the conversion of mandelonitrile and 2-phenylpropionitrile by a large set of variants generated from a nitrilase originating from pseudomonas fluorescens EBC191. *Molecules*, *24*(23). https://doi.org/10.3390/molecules24234232

Thuku, R. N., Brady, D., Benedik, M. J., & Sewell, B. T. (2009). Microbial nitrilases: Versatile, spiral forming, industrial enzymes. *Journal of Applied Microbiology*, *106*(3), 703–727. https://doi.org/10.1111/j.1365-2672.2008.03941.x

Todd, Z. R., & Öberg, K. I. (2020). Cometary Delivery of Hydrogen Cyanide to the Early Earth. *Astrobiology*, *20*(9), 1109–1120. https://doi.org/10.1089/ast.2019.2187

Veiga, M. M., Angeloci, G., Hitch, M., & Colon Velasquez-Lopez, P. (2014). Processing centres in artisanal gold mining. *Journal of Cleaner Production*, *64*, 535–544. https://doi.org/10.1016/j.jclepro.2013.08.015

Watanabe, A., Yano, K., Ikebukuro, K., & Karube, I. (1998). Cloning and expression of a gene encoding cyanidase from Pseudomonas stutzeri AK61. *Applied Microbiology and Biotechnology*, *50*(1), 93–97. https://doi.org/10.1007/s002530051261

CHAPTER 3. ANALYSIS OF THE DYNAMICS OF MUTATIONS IN SARS-CoV-2

3.1.- INTRODUCTION

SARS-CoV-2 is a virus that belongs to the order nidovirales and the coronaviridae family (Ben Hu et al., 2021). It is responsible for the COVID-19 pandemic declared by the World Health Organization on March 11th, 2020 (Cucinotta & Vanelli, 2020). Coronaviruses are RNA positive strand viruses with large genomes that infect mammals, birds and fishes. Coronaviruses are further classified into four groups: Alfa, Beta, Gamma and Delta. Two alphacoronavirus (HCoV-229E and HCoV-NL63) and two betacoronavirus (HCoV-OC43 and HKU-1) are known to cause common colds in humans (Chen et al., 2020). They are found in several animals in near contact with human populations and therefore are potential causes of zoonotic diseases. Two other betacoronavirus (different to SARS-CoV-2) are responsible for epidemics that originated Asia and the Middle-East; namely, SARS-CoV and MERS in 2002 and 2012, respectively (Fung & Liu, 2021). Those two betacoronaviruses, together with SARS-CoV-2, cause severe acute respiratory syndromes with variation in lethality and transmissibility of the disease (Fung & Liu, 2021).

Like other coronaviruses, SARS-CoV-2 presents an RNA genome with two regions: one that codes for non-structural proteins (nsps) and a second that codes for mainly structural proteins (Fig. 69) (Cui et al., 2019). In SARS-CoV-2, the first region codes for a polypeptide that, after the processing of two proteases (nsp3 and nsp5) forms sixteen independent peptides (nsp1 – nsp16) (see section 3.1.1. Life Cycle of SARS-CoV-2 for more details) (Ben Hu et al., 2021). Those proteins are responsible for the establishment of the replication-translation machinery of SARS-CoV-2 in the host cell. The second region in the genome codifies mainly structural proteins, such as the spike, envelope, membrane and nucleocapsid proteins that will form the SARS-CoV-2 virion. Furthermore, other proteins as the orf3a, orf6, orf7a, orf7b, orf8 and orf10 (named accessory proteins) are coded in this region (Cui et al., 2019; Fung & Liu, 2021; Ben Hu et al., 2021).

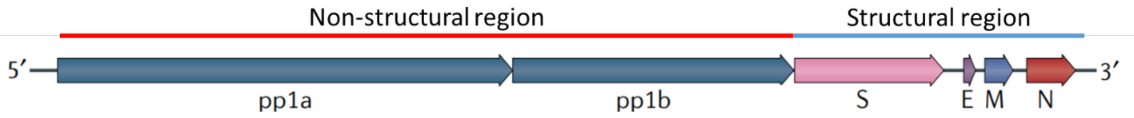


Figure 69. Scheme of the genome of SARS-CoV-2 showing the region of non-structural proteins and the region of structural proteins. (Figure adapted from. Cui et al. 2019)

3.1.1. Replication Cycle of SARS-CoV-2

The virion of SARS-CoV-2 is formed by the nucleocapsid-RNA complex wrapped in a membrane that presents three structural proteins: envelope (E), membrane (M) and spike (S) proteins (V'kovski et al., 2021). Spikes are trimeric proteins expressed on the surface of the virion and, after binding to the human Angiotensin converting enzyme-2 (ACE2) receptor, promotes the fusion of host and virion membranes (Li, 2016). Once inside the cell, the virion releases the RNA and translation of the non-structural proteins begins, pp1ab protein is processed by the protease domain of nsp3 releasing nsp1, nsp2 and nsp3. Nsp1 inhibits host-translation and degrades host-mRNA, whereas the function of nsp2 is unknown. The other protease from SARS-CoV-2, nsp5, releases the other nsp proteins including those involved in the double-membrane vesicle (DMV) formation (nsp4, nsp6), the replication complex (nsp7, nsp8, nsp12, nsp13 and nsp15) and the RNA processing machinery (nsp10, nsp14, nsp16) (Perlman & Netland, 2009). With the establishment of the replication-translation machinery complex, subgenomic mRNA for structural proteins are produced generating several copies of structural proteins that form new virions to be released by exocytosis (V'kovski et al., 2021). Accessory proteins such as orf3a are known to mediate immune evasion by different mechanisms such as interferon pathway inhibition (Kasuga et al., 2021).

3.1.2. Immunology of COVID-19

Until September 11, 2022, SARS-CoV-2 had caused 608 million registered infections and 6.51 million reported deaths all around the world (WHO. 2021). A large fraction of infected people recovers without the need for hospitalization (Nakamichi et al., 2021). However, approximately 18.4 % of infected people experience a disease that requires hospitalization and approximately half of these hospitalizations result in death (Nakamichi et al., 2021). The severe disease caused by SARS-CoV-2 is characterized by lymphopenia with a drastically reduction of CD4+, CD8+, B and NK cells. Also, reduction of monocytes, eosinophils and basophils is observed and an increase in neutrophils is often indicative of higher disease severity and poor outcome (Tan et al., 2020). Exhaustion markers normally expressed in chronic diseases (Wherry & Kurachi, 2015) are also upregulated in severe COVID-19 (Tan et al., 2020). Another feature in severe COVID-19 is the so-called cytokine storm where several types of mainly proinflammatory cytokines are released, such as IL-6, IL-1B, IL-2, IL-8, IL-17, GM-CSF, TNF (Melo et al., 2021). This cytokine storm may lead to shock and tissue damage in heart, liver, kidney and respiratory failure (Biying Hu et al., 2021).

3.1.3. Treatments and vaccines

Since the declaration of the pandemic by the WHO, several treatments and vaccines have been developed (Krammer, 2020). Despite the several studies performed all around the world, only two drugs have been shown to reduce the probability of death when administered. Dexamethasone is a glucocorticoid that modulates inflammation. A randomized trial showed that treatment with dexamethasone in patients receiving invasive mechanical ventilation treatment resulted in reduced fatalities when compared to those not treated with dexamethasone (29.3 % vs. 41.4 %) (The RECOVERY collaborative group, 2021). The differences were less when patients receiving oxygen but without invasive mechanical ventilation were treated with dexamethasone (23.3 % vs 26.2 %) and greater incidence of death was observed when patients did not receive respiratory support (17.8 % vs. 14.0 %) (The RECOVERY collaborative group, 2021). Additionally, the anti-interleukin-6 receptor antibody tocilizumab resulted in a reduced proportion of patients requiring mechanical ventilation or who died after 28 days of treatment with this antibody (12 % vs. 19.3 %) (Salama et al., 2021). On the other hand, patients with severe COVID-19 treated with tocilizumab did not show better clinical status or lower mortality than placebo (Rosas et al., 2021). Also, tocilizumab was not effective at preventing intubation or death in moderately ill hospitalized patients (Stone et al., 2020).

The most promising method to bring the pandemic to an end are vaccines. This preventive treatment has been useful for other infectious diseases in human history (Pollard & Bijker, 2021). There was an unprecedented international effort to produce effective and safe vaccines, with roughly 180 vaccines being developed around the world during 2020 (Krammer, 2020).

There are at least three types of vaccines being developed, with several of them already authorized in many countries (Kim et al., 2021). Whole-virus vaccines can be produced by rationally weakening the virus using mutational methodologies, cell-culture passages, or inactivation (i.e: thermally or chemically denatured whole-virus) (Kyriakidis et al., 2021). The advantage of these vaccines is that the whole-virus is presented to the host-immunological system. However, this type of vaccine requires biosafety level 3 laboratories for production. That means that the production could be slow due to the few laboratories that exist with these requirements and the need to use cell-cultures for the replication of the virus (Dong et al., 2020).

Coronavac (Sinovac) and Sinopharm are two vaccines that use SARS-CoV-2 inactivated with B-propiolactone (Gao et al., 2020; H. Wang et al., 2020). Phase I/II clinical trials showed safety and induction of neutralizing antibodies that allowed for their emergency authorization (Wu et al., 2021; Xia et al., 2020; Y. Zhang et al., 2021) in a regimen of two doses 21 days apart for the Coronavac vaccine or 14 days apart for the Sinopharm vaccine. The only phase III clinical trial for CoronaVac published in a peer-reviewed journal showed 83.5 % overall vaccine efficacy (Tanriover et al., 2021).

Viral-vector vaccines use a template virus to superficially express the target protein (i.e spike protein from SARS-CoV-2). The template virus could be a replicating or non-replicating virus. Normally, this vaccine expresses just one protein from the pathogen; thus, generating an immune response against just one part of the virus. However, the presence of proteins from the template virus functions as an adjuvant to stimulate a stronger immune response (Iwasaki & Omer, 2020; Tatsis & Ertl, 2004; C. Zhang & Zhou, 2016).

Sputnik V, ChAdOx1-nCoV and Janssen vaccines use this technology but with different viral vectors expressing the spike SARS-CoV-2 protein (Bos et al., 2020; Logunov et al., 2020; van Doremalen et al., 2020). Sputnik V uses two different adenovirus Ad-26 and Ad-5 (Logunov et al., 2020), ChAdOx1-nCoV uses a chimpanzee adenovirus (ChAdOx1) (van Doremalen et al., 2020) and Janssen vaccine uses Ad-26 (Bos et al., 2020). These three vaccines have already published the results of their phase III clinical trials showing safety and efficacy sufficient to be considered as new therapies to control COVID-19 (Logunov et al., 2021, Voysey et al., 2021, Sadoff et al., 2021).

The Sputnik V vaccine showed an overall efficacy of 91.6 % with very similar results in the different age groups tested. The regimen tested for this vaccine was two doses 21 days apart (Logunov et al., 2021). On the other hand, the ChAdOx1-nCov vaccine showed less overall efficacy (70.4 %); although, in this case several regimens were tested. The regimen consisting in a low dose (2.2 x 1010 viral particles) followed by a standard dose (~5 x 1010 viral particles) resulted in a calculated efficacy of 90 %. Other regimens tested for this vaccine had a calculated efficacy between 60 and 75 % approximately (Voysey et al., 2021). For the Janssen vaccine, phase III clinical trials showed an overall efficacy of 66.9 % with a slightly higher efficacy against severe-critical COVID-19 of 76.7 % (Sadoff et al., 2021). The difference between this vaccine and the others is that the tested regimen was of a single dose (Sadoff et al., 2021).

Nucleic-acid vaccines are a relatively new approach to deliver a gene or RNA message coding for one or more target proteins to a host-cell to induce the production of pathogen-proteins in the membrane and allow the immune system to respond against this (Gary & Weiner, 2020; Pardi et al., 2018, 2020). Two RNA-based vaccines have published their phase III clinical trials: Moderna and Pfizer (Baden et al., 2021; Polack et al., 2020). These two vaccines showed an overall efficacy of more than 94 % that varied very little among age groups. To obtain these efficacies, the regimen tested for the Moderna vaccine was two doses of 100 µg mRNA-1273, 28 days apart (Baden et al., 2021). In the case of Pfizer regimen, two doses of 30 ug of BNT162b2 vaccine was administered 21 days apart (Polack et al., 2020).

3.1.4. Variants

Since the beginning of the COVID-19 pandemic, several SARS-CoV-2 variants were described. Early in the pandemic, the mutation D614G in the spike protein was reported in Europe and a relatively high increase in its frequency of isolation was hypothesized due to an increase in the transmissibility of SARS-CoV-2 (Korber et al., 2020). Residue 614 of the spike protein is in the interface between the monomers of the trimers and it was hypothesized that the effect in the stability of the trimer could affect the capacity of fusion between the SARS-CoV-2 virion and the host-cell (Gobeil et al., 2021; Zhang et al., 2021). Epidemiological studies based in phylogenetic approaches showed not very confident results to propose an increment of the transmissibility due to a direct effect of D614G mutation, with similar probabilities of being caused by a random (i.e: genetic drift) effect (Volz et al., 2021). Also, most of the SARS-CoV-2 genomes (more than 99 %

of them) present three other mutations compared to the reference genome: P323L in the nsp12 protein, C241T in the 5´UTR and C3037T (synonymous mutation) in the nsp3 protein (van Dorp et al., 2020). Currently, the only *in vivo* experimental evidence (not in humans) indicating that D614G may result in a slight increase in the transmission capacity is that it shows a more rapid propagation between hamsters in contiguous cages (Plante et al., 2021).

More recently, the effect in transmissibility, pathogenicity, and the effect in the efficacy of vaccines of three lineages has been discussed (Harvey et al., 2021). Lineages B.1.1.7 (alpha), B.1.351 (beta) and P.1 (gamma) appeared apparently independent in different parts of the world with some common mutations (Chaillon & Smith, 2021; Faria et al., 2021; Tegally et al., 2021). Two of the most investigated are the N501Y (present in the three lineages) and E484K (present in B.1.351 and P.1) in the spike protein affects the binding to the receptor ACE2 (Jangra et al., 2021; Tian et al., 2021; Yang et al., 2021; Zhu et al., 2021). N501Y is present in the RBD domain near to the interaction site with ACE2. However, slight or no considerable reduction in the neutralizing capacity of antibodies generated by recovered or vaccinated people has been observed (Bates et al., 2021; Collier et al., 2021; Plante et al., 2021; Supasa et al., 2021; Xie et al., 2021). E484K is also localized in the RBD domain, but not in the region that directly interacts with ACE2 (Gobeil et al., 2021). E484K was also postulated as a probable candidate to promote immune-escape; in the case of this mutant, studies showed that the capacity of convalescent sera or vaccinated sera (from Pfizer or Moderna vaccines) has slightly less neutralization potential compared to the wild-type (Jangra et al., 2021; Z. Wang et al., 2021; Xie et al., 2021). In a recent report, the combination of the mutations presents in lineage B.1.1.7 with the mutation E484K showed a greater reduction of the sera from vaccinated people (with Moderna vaccine) to neutralize the SARS-CoV-2 (Collier et al., 2021).

The last two SARS-CoV-2 variants that raised concerns were Delta and Omicron. Delta presents eight mutations in the Spike protein respect to the first SARS-CoV-2 genome reported (T19R, del156-157, R158G, L452R, T478K, D614G, P681R, and D950N). Delta was able to replace the preexisting lineages in several countries where it was reported (Earnest et al., 2022, Toole et al., 2021), likely due to its higher transmissibility and its capacity to better evade the immune responses elicited by vaccination (Earnest et al., 2022, Farinholt et al., 2021, Zhang et al., 2021).

Since November 2021, the Omicron variant first reported in Africa have reached several countries causing the highest number of COVID-19 cases seen since the beginning of the pandemic (Susuki et al. 2022, Vianna et al. 2022). This variant has 33 mutations in the Spike protein respect to the first reported genome of SARS-CoV-2 (A67V, del69-70, T95I, del142-144, Y145D, del211, L212I, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, and L981F). Omicron variant replicates faster than other variants in the bronchi but less efficiently in the lung parenchyma and its spike is less fusogenic (Susuki et al. 2022, Hui et al. 2022). This can explain at least in part the less severity reported for this variant (Susuki et al. 2022). Additionally, it was shown that vaccines effectiveness and neutralization by convalescent-vaccinated people against Omicron is reduced, but still protects against severe and critical illness (Collie et al. 2022, Rossler et al. 2022, Altarawneh et al. 2022, Liu et al. 2021).

Since the beggining of the pandemic, the monitoring of the appearance of mutations and new variants of concern was established. Two articles were published based on these analyses. The first reports the determination and spatiotemporal analysis of five major haplotypes and the second analyzes the regional specificity of several SARS-CoV-2 mutations and the possible impact of control measures on the mutation of concern N501Y.

3.2.- OBJECTIVES

The results of this chapter are presented in two published articles attached to this thesis:

1.- Santiago Justo Arevalo, Daniela Zapata Sifuentes, Cesar J. Huallpa, Gianfranco Landa Bianchi, Adriana Castillo Chavez, Romina Garavito-Salini Casas, Carmen Sofia Uribe Calampa, Guillermo Uceda-Campos, Roberto Pineda Chavarria. Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures. Scientific reports 11, 17755 (2021). https://doi.org/10.1038/s41598-021-97267-7. (Appendix 3.1).

2.- Santiago Justo Arevalo, Daniela Zapata Sifuentes, Cesar J. Huallpa, Gianfranco Landa Bianchi, Adriana Castillo Chavez, Romina Garavito-Salini Casas, Guillermo Uceda-Campos, Roberto Pineda Chavarria. Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes normalized by COVID-19 cases during the pandemic. Frontiers in

Microbiology 12, 612432 (2021). https://doi.org/10.3389/fmicb.2021.612432. (Appendix 3.2).

The main aim of these studies was to perform a geographical and temporal analysis of emergent mutations in the SARS-CoV-2 genome up to April 2021. The following specific objectives were achieved during the process:

- Determination of SARS-CoV-2 major haplotypes.
- Analysis of the geographical and temporal patterns of distribution of the major haplotypes.
- Identification of mutations of concern based on the estimation of the number of cases where they are present.
- Analysis of the temporal dynamics of the mutations of concern.
- Analysis of the regional specificity of the mutations of concern.
- Correlation analysis of mutations with the level of control measures.

3.3.- REFERENCES

Altarawneh, H., Chemaitelly, H., Hasan, M., Ayoub, H., Qassim, S., AlMukdad, S., Coyle, P., Yassine, H., Al-Khatib, H., Benslimane, F., Al-Kanaani, Z., Al-Kuwari, E., Jeremijenko, A., Kaleeckal, A., Latif, A., Shaik, R., Abdul-Rahim, H., Nasrallah, G., Al-Kuwari, M., … Abu-Taddad, L. (2022). Protection against the Omicron Variant from Previous SARS-CoV-2 Infection. *New England Journal of Medicine*, *386*(13).

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., … Zaks, T. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*, *384*(5), 403–416. https://doi.org/10.1056/nejmoa2035389

Bates, T. A., Leier, H. C., Lyski, Z. L., McBride, S. K., Coulter, F. J., Weinstein, J. B., Goodman, J. R., Lu, Z., Siegel, S. A. R., Sullivan, P., Strnad, M., Brunton, A. E., Lee, D. X., Adey, A. C., Bimber, B. N., O'Roak, B. J., Curlin, M. E., Messer, W. B., & Tafesse, F. G. (2021). Neutralization of SARS-CoV-2 variants by convalescent and BNT162b2 vaccinated serum. *Nature Communications*, *12*(1), 5135. https://doi.org/10.1038/s41467-021-25479-6

Bos, R., Rutten, L., van der Lubbe, J. E. M., Bakkers, M. J. G., Hardenberg, G., Wegmann, F., Zuijdgeest, D., de Wilde, A. H., Koornneef, A., Verwilligen, A., van Manen, D., Kwaks, T., Vogels, R., Dalebout, T. J., Myeni, S. K., Kikkert, M., Snijder, E. J., Li, Z., Barouch, D. H., … Schuitemaker, H. (2020). Ad26 vector-based COVID-19 vaccine encoding a prefusion-stabilized SARS-CoV-2 Spike immunogen induces potent humoral and cellular immune responses. *Npj Vaccines*, *5*(1), 1–11. https://doi.org/10.1038/s41541-020-00243-x

Chaillon, A., & Smith, D. M. (2021). Phylogenetic Analyses of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) B.1.1.7 Lineage Suggest a Single Origin Followed by Multiple Exportation Events Versus Convergent Evolution. *Clinical Infectious Diseases*, *2*(Xx Xxxx), 5–8. https://doi.org/10.1093/cid/ciab265

Chen, Y., Liu, Q., & Guo, D. (2020). Emerging coronaviruses: Genome structure, replication, and pathogenesis. *Journal of Medical Virology*, *92*(4), 418–423.

https://doi.org/10.1002/jmv.25681

Collie, S., Champion, J., Moultrie, H., Bekker, L.-G., & Gray, G. (2022). Third BNT162b2 Vaccination Neutralization of SARS-CoV-2 Omicron Infection. *New England Journal of Medicine*, *386*(5), 492–494. https://doi.org/10.1056/nejmc2119358

Collier, D. A., De Marco, A., Ferreira, I. A. T. M., Meng, B., Datir, R. P., Walls, A. C., Kemp, S. A., Bassi, J., Pinto, D., Silacci-Fregni, C., Bianchi, S., Tortorici, M. A., Bowen, J., Culap, K., Jaconi, S., Cameroni, E., Snell, G., Pizzuto, M. S., Pellanda, A. F., … Gupta, R. K. (2021). Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature*, *593*(7857), 136–141. https://doi.org/10.1038/s41586-021-03412-7

Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Biomedica*, *91*(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, *17*(3), 181–192. https://doi.org/10.1038/s41579-018-0118-9

Dong, Y., Dai, T., Wei, Y., Zhang, L., Zheng, M., & Zhou, F. (2020). A systematic review of SARS-CoV-2 vaccine candidates. *Signal Transduction and Targeted Therapy*, *5*(1). https://doi.org/10.1038/s41392-020-00352-y

Earnest, R., Uddin, R., Matluk, N., Renzette, N., Turbett, S. E., Siddle, K. J., Loreth, C., Adams, G., Tomkins-Tinch, C. H., Petrone, M. E., Rothman, J. E., Breban, M. I., Koch, R. T., Billig, K., Fauver, J. R., Vogels, C. B. F., Bilguvar, K., De Kumar, B., Landry, M. L., … Mandese, Z. M. (2022). Comparative transmissibility of SARS-CoV-2 variants

Delta and Alpha in New England, USA. *Cell Reports Medicine*, 100583. https://doi.org/10.1016/j.xcrm.2022.100583

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C. S., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., … Sabino, E. C. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science (New York, N.Y.)*, *372*(6544), 815–821. https://doi.org/10.1126/science.abh2644

Farinholt, T., Doddapaneni, H., Qin, X., Menon, V., Meng, Q., Metcalf, G., Chao, H., Gingras, M. C., Avadhanula, V., Farinholt, P., Agrawal, C., Muzny, D. M., Piedra, P. A., Gibbs, R. A., & Petrosino, J. (2021). Transmission event of SARS-CoV-2 delta variant reveals multiple vaccine breakthrough infections. *BMC Medicine*, *19*(1), 1–6. https://doi.org/10.1186/s12916-021-02103-4

Fung, T. S., & Liu, D. X. (2021). Similarities and Dissimilarities of COVID-19 and Other Coronavirus Diseases. *Annual Review of Microbiology*, *75*(1), 1–29. https://doi.org/10.1146/annurev-micro-110520-023212

Gao, Q., Bao, L., Mao, H., Wang, L., Xu, K., Yang, M., Li, Y., Zhu, L., Wang, N., Lv, Z., Gao, H., Ge, X., Kan, B., Hu, Y., Liu, J., Cai, F., Jiang, D., Yin, Y., Qin, C., … Qin, C. (2020). Development of an inactivated vaccine candidate for SARS-CoV-2. *Science*, *369*(6499), 77–81. https://doi.org/10.1126/science.abc1932

Gary, E. N., & Weiner, D. B. (2020). DNA vaccines: prime time is now. *Current Opinion in*

Immunology, *65*, 21–27. https://doi.org/10.1016/j.coi.2020.01.006

Gobeil, S. M. C., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Stalls, V., Kopp, M. F., Manne, K., Li, D., Wiehe, K., Saunders, K. O., Edwards, R. J., Korber, B., Haynes, B. F., Henderson, R., & Acharya, P. (2021). Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science*, *373*(6555), eabi6226. https://doi.org/10.1126/science.abi6226

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), 409–424. https://doi.org/10.1038/s41579-021-00573-0

Hu, Ben, Guo, H., Zhou, P., & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7

Hu, Biying, Huang, S., & Yin, L. (2021). The cytokine storm and COVID-19. *Journal of Medical Virology*, *93*(1), 250–256. https://doi.org/10.1002/jmv.26232

Hui, K. P. Y., Ho, J. C. W., Cheung, M. chun, Ng, K. chun, Ching, R. H. H., Lai, K. ling, Kam, T. T., Gu, H., Sit, K. Y., Hsin, M. K. Y., Au, T. W. K., Poon, L. L. M., Peiris, M., Nicholls, J. M., & Chan, M. C. W. (2022). SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature*, *603*(7902), 715–720. https://doi.org/10.1038/s41586-022-04479-6

Iwasaki, A., & Omer, S. B. (2020). Why and How Vaccines Work. *Cell*, *183*(2), 290–295.

https://doi.org/10.1016/j.cell.2020.09.040

Jangra, S., Ye, C., Rathnasinghe, R., Stadlbauer, D., Alshammary, H., Amoako, A. A., Awawda, M. H., Beach, K. F., Bermúdez-González, M. C., Chernet, R. L., Eaker, L. Q., Ferreri, E. D., Floda, D. L., Gleason, C. R., Kleiner, G., Jurczyszak, D., Matthews, J. C., Mendez, W. A., Mulder, L. C. F., … Schotsaert, M. (2021). SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *The Lancet Microbe*, *2*(7), e283–e284. https://doi.org/10.1016/S2666-5247(21)00068-9

Kasuga, Y., Zhu, B., Jang, K. J., & Yoo, J. S. (2021). Innate immune sensing of coronavirus and viral evasion strategies. *Experimental and Molecular Medicine*, *53*(5), 723–736. https://doi.org/10.1038/s12276-021-00602-1

Kim, J. H., Marks, F., & Clemens, J. D. (2021). Looking beyond COVID-19 vaccine phase 3 trials. *Nature Medicine*, *27*(2), 205–211. https://doi.org/10.1038/s41591-021-01230-y

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., Angyal, A., Brown, R. L., Carrilero, L., … Montefiori, D. C. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*, *182*(4), 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043

Krammer, F. (2020). SARS-CoV-2 vaccines in development. *Nature*, *586*(7830), 516–527. https://doi.org/10.1038/s41586-020-2798-3

Kyriakidis, N. C., López-Cortés, A., González, E. V., Grimaldos, A. B., & Prado, E. O. (2021). SARS-CoV-2 vaccines strategies: a comprehensive review of phase 3 candidates. *Npj Vaccines*, *6*(1). https://doi.org/10.1038/s41541-021-00292-w

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology*, *3*, 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Liu, L., Iketani, S., Guo, Y., Chan, J. F. W., Wang, M., Liu, L., Luo, Y., Chu, H., Huang, Y., Nair, M. S., Yu, J., Chik, K. K. H., Yuen, T. T. T., Yoon, C., To, K. K. W., Chen, H., Yin, M. T., Sobieszczyk, M. E., Huang, Y., … Ho, D. D. (2022). Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature*, *602*(7898), 676–681. https://doi.org/10.1038/s41586-021-04388-0

Logunov, D. Y., Dolzhikova, I. V., Shcheblyakov, D. V., Tukhvatulin, A. I., Zubkova, O. V., Dzharullaeva, A. S., Kovyrshina, A. V., Lubenets, N. L., Grousova, D. M., Erokhova, A. S., Botikov, A. G., Izhaeva, F. M., Popova, O., Ozharovskaya, T. A., Esmagambetov, I. B., Favorskaya, I. A., Zrelkin, D. I., Voronina, D. V., Shcherbinin, D. N., … Gintsburg, A. L. (2021). Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia. *The Lancet*, *397*(10275), 671–681. https://doi.org/10.1016/S0140-6736(21)00234-8

Logunov, D. Y., Dolzhikova, I. V., Zubkova, O. V., Tukhvatullin, A. I., Shcheblyakov, D. V., Dzharullaeva, A. S., Grousova, D. M., Erokhova, A. S., Kovyrshina, A. V., Botikov, A. G., Izhaeva, F. M., Popova, O., Ozharovskaya, T. A., Esmagambetov, I. B., Favorskaya, I. A., Zrelkin, D. I., Voronina, D. V., Shcherbinin, D. N., Semikhin, A. S., … Gintsburg, A. L. (2020). Safety and immunogenicity of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine in two formulations: two open, non-randomised phase 1/2 studies from Russia. *The Lancet*, *396*(10255), 887–897. https://doi.org/10.1016/S0140-6736(20)31866-3

Melo, A. K. G., Milby, K. M., Caparroz, A. L. M. A., Pinto, A. C. P. N., Santos, R. R. P., Rocha, A. P., Ferreira, G. A., Souza, V. A., Valadares, L. D. A., Vieira, R. M. R. A., Pileggi, G. S., & Trevisani, V. F. M. (2021). Biomarkers of cytokine storm as red flags for severe and fatal COVID-19 cases: A living systematic review and meta-analysis. *PLoS ONE*, *16*(6 June), 1–21. https://doi.org/10.1371/journal.pone.0253894

Nakamichi, K., Shen, J. Z., Lee, C. S., Lee, A., Roberts, E. A., Simonson, P. D., Roychoudhury, P., Andriesen, J., Randhawa, A. K., Mathias, P. C., Greninger, A. L., Jerome, K. R., & Van Gelder, R. N. (2021). Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-82850-9

Pardi, N., Hogan, M. J., Porter, F. W., & Weissman, D. (2018). mRNA vaccines-a new era in vaccinology. *Nature Reviews Drug Discovery*, *17*(4), 261–279. https://doi.org/10.1038/nrd.2017.243

Pardi, N., Hogan, M. J., & Weissman, D. (2020). Recent advances in mRNA vaccine technology. *Current Opinion in Immunology*, *65*, 14–20. https://doi.org/10.1016/j.coi.2020.01.008

Perlman, S., & Netland, J. (2009). Coronaviruses post-SARS: Update on replication and pathogenesis. *Nature Reviews Microbiology*, *7*(6), 439–450. https://doi.org/10.1038/nrmicro2147

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., … Shi, P. Y. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, *592*(7852), 116–121. https://doi.org/10.1038/s41586-020-2895-3

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. https://doi.org/10.1056/nejmoa2034577

Pollard, A. J., & Bijker, E. M. (2021). A guide to vaccinology: from basic principles to new developments. *Nature Reviews Immunology*, *21*(2), 83–100. https://doi.org/10.1038/s41577-020-00479-7

Rosas, I. O., Bräu, N., Waters, M., Go, R. C., Hunter, B. D., Bhagani, S., Skiest, D., Aziz, M. S., Cooper, N., Douglas, I. S., Savic, S., Youngstein, T., Del Sorbo, L., Cubillo Gracian, A., De La Zerda, D. J., Ustianowski, A., Bao, M., Dimonaco, S., Graham, E., … Malhotra, A. (2021). Tocilizumab in Hospitalized Patients with Severe Covid-19 Pneumonia. *New England Journal of Medicine*, *384*(16), 1503–1516.

https://doi.org/10.1056/nejmoa2028700

Rossler, A., Riepler, L., Bante, D., von Laer, D., & Kimpel, J. (2022). SARS-CoV-2 Omicron Variant Neutralization in Serum from Vaccinated and Convalescent Persons. *New England Journal of Medicine*, *386*(7), 1–4.

Sadoff, J., Gray, G., Vandebosch, A., Cárdenas, V., Shukarev, G., Grinsztejn, B., Goepfert, P. A., Truyers, C., Fennema, H., Spiessens, B., Offergeld, K., Scheper, G., Taylor, K. L., Robb, M. L., Treanor, J., Barouch, D. H., Stoddard, J., Ryser, M. F., Marovich, M. A., … Douoguih, M. (2021). Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *New England Journal of Medicine*, *384*(23), 2187–2201. https://doi.org/10.1056/nejmoa2101544

Salama, C., Han, J., Yau, L., Reiss, W. G., Kramer, B., Neidhart, J. D., Criner, G. J., Kaplan-Lewis, E., Baden, R., Pandit, L., Cameron, M. L., Garcia-Diaz, J., Chávez, V., Mekebeb-Reuter, M., Lima de Menezes, F., Shah, R., González-Lara, M. F., Assman, B., Freedman, J., & Mohan, S. V. (2021). Tocilizumab in Patients Hospitalized with Covid-19 Pneumonia. *New England Journal of Medicine*, *384*(1), 20–30. https://doi.org/10.1056/nejmoa2030340

Stone, J. H., Frigault, M. J., Serling-Boyd, N. J., Fernandes, A. D., Harvey, L., Foulkes, A. S., Horick, N. K., Healy, B. C., Shah, R., Bensaci, A. M., Woolley, A. E., Nikiforow, S., Lin, N., Sagar, M., Schrager, H., Huckins, D. S., Axelrod, M., Pincus, M. D., Fleisher, J., … Mansour, M. K. (2020). Efficacy of Tocilizumab in Patients Hospitalized with Covid-19. *New England Journal of Medicine*, *383*(24), 2333–2344. https://doi.org/10.1056/nejmoa2028836

Supasa, P., Zhou, D., Dejnirattisai, W., Liu, C., Mentzer, A. J., Ginn, H. M., Zhao, Y., Duyvesteyn, H. M. E., Nutalai, R., Tuekprakhon, A., Wang, B., Paesen, G. C., Slon-Campos, J., López-Camacho, C., Hallis, B., Coombes, N., Bewley, K. R., Charlton, S., Walter, T. S., … Screaton, G. R. (2021). Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell*, *184*(8), 2201-2211.e7. https://doi.org/10.1016/j.cell.2021.02.033

Suzuki, R., Yamasoba, D., Kimura, I., Wang, L., Kishimoto, M., Ito, J., Morioka, Y., Nao, N., Nasser, H., Uriu, K., Kosugi, Y., Tsuda, M., Orba, Y., Sasaki, M., Shimizu, R., Kawabata, R., Yoshimatsu, K., Asakura, H., Nagashima, M., … Sato, K. (2022). Attenuated fusogenicity and pathogenicity of SARS-CoV-2 Omicron variant. *Nature*, *603*(December 2021). https://doi.org/10.1038/s41586-022-04462-1

Tan, M., Liu, Y., Zhou, R., Deng, X., Li, F., Liang, K., & Shi, Y. (2020). Immunopathological characteristics of coronavirus disease 2019 cases in Guangzhou, China. *Immunology*, *160*(3), 261–268. https://doi.org/10.1111/imm.13223

Tanriover, M. D., Doğanay, H. L., Akova, M., Güner, H. R., Azap, A., Akhan, S., Köse, Ş., Erdinç, F. Ş., Akalın, E. H., Tabak, Ö. F., Pullukçu, H., Batum, Ö., Şimşek Yavuz, S., Turhan, Ö., Yıldırmak, M. T., Köksal, İ., Taşova, Y., Korten, V., Yılmaz, G., … Aksu, K. (2021). Efficacy and safety of an inactivated whole-virion SARS-CoV-2 vaccine (CoronaVac): interim results of a double-blind, randomised, placebo-controlled, phase 3 trial in Turkey. *The Lancet*, *398*(10296), 213–222. https://doi.org/10.1016/S0140-6736(21)01429-X

Tatsis, N., & Ertl, H. C. J. (2004). Adenoviruses as vaccine vectors. *Molecular Therapy*, *10*(4), 616–629. https://doi.org/10.1016/j.ymthe.2004.07.013

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., … de Oliveira, T. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, *592*(7854), 438–443. https://doi.org/10.1038/s41586-021-03402-9

The RECOVERY collaborative group. (2021). Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine*, *384*(8), 693–704. https://doi.org/10.1056/nejmoa2021436

Tian, F., Tong, B., Sun, L., Shi, S., Zheng, B., Wang, Z., Dong, X., & Zheng, P. (2021). N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *ELife*, *10*, 1–17. https://doi.org/10.7554/elife.69091

Toole, Á. O., Hill, V., Pybus, O. G., Watts, A., Bogoch, I. I., Khan, K., Messina, J. P., Network, B. C. G., Tegally, H., Lessells, R. R., Giandhari, J., Pillay, S., Tumedi, K. A., Merhi, G., Koweyes, J., Geoghegan, J. L., Ligt, J. De, Ren, X., Storey, M., … Munnink, B. O. (2021). Tracking the international spread of SARS-CoV-2 lineages B . 1 . 1 . 7 and B . 1 . 351 / 501Y-V2. *Wellcome Open Research*, *6*(121), 1–14.

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology*, *19*(3), 155–170. https://doi.org/10.1038/s41579-020-00468-6

van Doremalen, N., Lambe, T., Spencer, A., Belij-Rammerstorfer, S., Purushotham, J. N., Port, J. R., Avanzato, V. A., Bushmaker, T., Flaxman, A., Ulaszewska, M., Feldmann, F., Allen, E. R., Sharpe, H., Schulz, J., Holbrook, M., Okumura, A., Meade-White, K., Pérez-Pérez, L., Edwards, N. J., … Munster, V. J. (2020). ChAdOx1 nCoV-19 vaccine prevents SARS-CoV-2 pneumonia in rhesus macaques. *Nature*, *586*(7830), 578–582. https://doi.org/10.1038/s41586-020-2608-y

van Dorp, L., Richard, D., Tan, C. C. S., Shaw, L. P., Acman, M., & Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19818-2

Viana, R., Moyo, S., Amoako, D. G., Tegally, H., Scheepers, C., Althaus, C. L., Anyaneji, U. J., Bester, P. A., Boni, M. F., Chand, M., Choga, W. T., Colquhoun, R., Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., … de Oliveira, T. (2022). Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, *603*(7902), 679–686. https://doi.org/10.1038/s41586-022-04411-y

Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., Rey, S. M., Nicholls, S. M., Colquhoun, R. M., da Silva Filipe, A., Shepherd, J., Pascall, D. J., Shah, R., Jesudason, N., Li, K., … Pybus, O. G. (2021). Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*, *184*(1), 64-75.e11. https://doi.org/10.1016/j.cell.2020.11.020

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., Angus, B., Baillie, V. L., Barnabas, S. L., Bhorat, Q. E., Bibi, S., Briner, C., Cicconi, P., Collins, A. M., Colin-Jones, R., Cutland, C. L., Darton, T. C., Dheda, K., Duncan, C. J. A., … Zuidewind, P. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet*, *397*(10269), 99–111. https://doi.org/10.1016/S0140-6736(20)32661-1

Wang, H., Zhang, Y., Huang, B., Deng, W., Quan, Y., Wang, W., Xu, W., Zhao, Y., Li, N., Zhang, J., Liang, H., Bao, L., Xu, Y., Ding, L., Zhou, W., Gao, H., Liu, J., Niu, P., Zhao, L., … Yang, X. (2020). Development of an Inactivated Vaccine Candidate, BBIBP-CorV, with Potent Protection against SARS-CoV-2. *Cell*, *182*(3), 713-721.e9. https://doi.org/10.1016/j.cell.2020.06.008

Wang, Z., Schmidt, F., Weisblum, Y., Muecksch, F., Barnes, C. O., Finkin, S., Schaefer-Babajew, D., Cipolla, M., Gaebler, C., Lieberman, J. A., Oliveira, T. Y., Yang, Z., Abernathy, M. E., Huey-Tubman, K. E., Hurley, A., Turroja, M., West, K. A., Gordon, K., Millard, K. G., … Nussenzweig, M. C. (2021). mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature*, *592*(7855), 616–622. https://doi.org/10.1038/s41586-021-03324-6

Wherry, E. J., & Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nature Reviews Immunology*, *15*(8), 486–499. https://doi.org/10.1038/nri3862

Wu, Z., Hu, Y., Xu, M., Chen, Z., Yang, W., Jiang, Z., Li, M., Jin, H., Cui, G., Chen, P., Wang, L., Zhao, G., Ding, Y., Zhao, Y., & Yin, W. (2021). Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine (CoronaVac) in healthy adults aged 60 years and older: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *The Lancet Infectious Diseases*, *21*(6), 803–812. https://doi.org/10.1016/S1473-3099(20)30987-7

Xia, S., Duan, K., Zhang, Y., Zhao, D., Zhang, H., Xie, Z., Li, X., Peng, C., Zhang, Y., Zhang, W., Yang, Y., Chen, W., Gao, X., You, W., Wang, X., Wang, Z., Shi, Z., Wang, Y., Yang, X., … Yang, X. (2020). Effect of an Inactivated Vaccine Against SARS-CoV-2 on Safety and Immunogenicity Outcomes: Interim Analysis of 2 Randomized Clinical Trials. *JAMA - Journal of the American Medical Association*, *324*(10), 951–960. https://doi.org/10.1001/jama.2020.15543

Xie, X., Liu, Y., Liu, J., Zhang, X., Zou, J., Fontes-Garfias, C. R., Xia, H., Swanson, K. A., Cutler, M., Cooper, D., Menachery, V. D., Weaver, S. C., Dormitzer, P. R., & Shi, P. Y. (2021). Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nature Medicine*, *27*(4), 620–621. https://doi.org/10.1038/s41591-021-01270-4

Yang, T. J., Yu, P. Y., Chang, Y. C., Liang, K. H., Tso, H. C., Ho, M. R., Chen, W. Y., Lin, H. T., Wu, H. C., & Hsu, S. T. D. (2021). Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. *Nature Structural and Molecular Biology*, *28*(SePtember). https://doi.org/10.1038/s41594-021-00652-z

Zhang, C., & Zhou, D. (2016). Adenoviral vector-based strategies against infectious disease and cancer. *Human Vaccines and Immunotherapeutics*, *12*(8), 2064–2074. https://doi.org/10.1080/21645515.2016.1165908

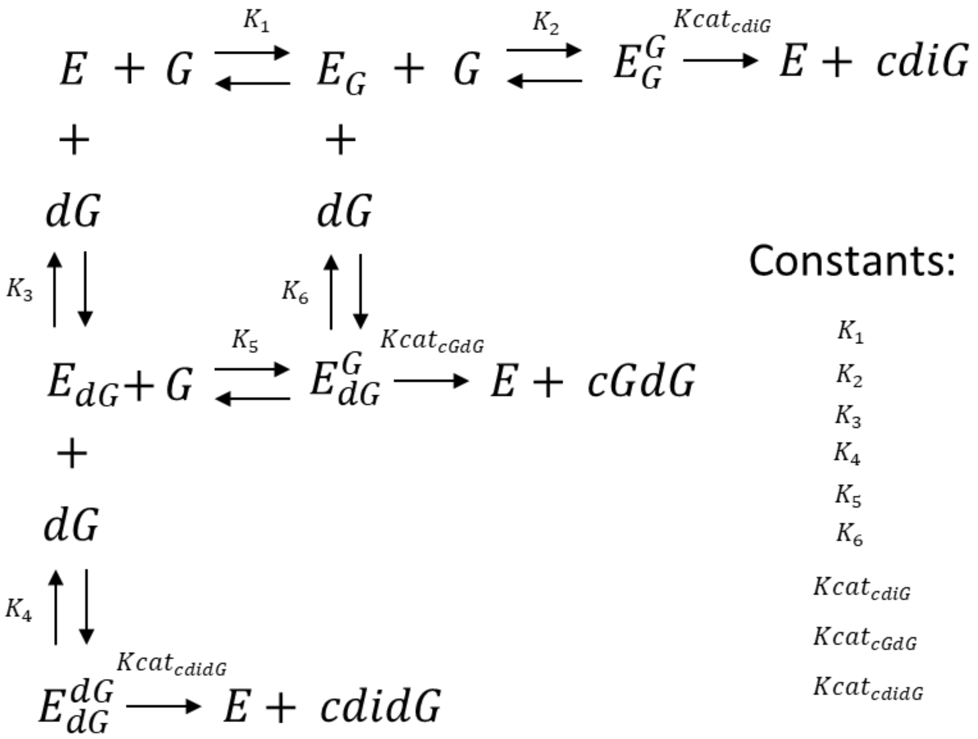
Zhang, J., Cai, Y., Xiao, T., Lu, J., Peng, H., Sterling, S. M., Walsh, R. M., Rits-Volloch, S., Zhu, H., Woosley, A. N., Yang, W., Sliz, P., & Chen, B. (2021). Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science*, *372*(6541), 525–530. https://doi.org/10.1126/science.abf2303

Zhang, J., Xiao, T., Cai, Y., Lavine, C. L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M. L., Walsh, R. M., Rits-Volloch, S., Wesemann, D. R., Yang, W., Seaman, M. S., Lu, J., & Chen, B. (2021). Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. *Science*, *374*(6573), 1353–1360. https://doi.org/10.1126/science.abl9463

Zhang, Y., Zeng, G., Pan, H., Li, C., Hu, Y., Chu, K., Han, W., Chen, Z., Tang, R., Yin, W., Chen, X., Hu, Y., Liu, X., Jiang, C., Li, J., Yang, M., Song, Y., Wang, X., Gao, Q., & Zhu, F. (2021). Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine in healthy adults aged 18–59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *The Lancet Infectious Diseases*, *21*(2), 181–192. https://doi.org/10.1016/S1473-3099(20)30843-4

Zhu, X., Mannar, D., Srivastava, S. S., Berezuk, A. M., Demers, J. P., Saville, J. W., Leopold, K., Li, W., Dimitrov, D. S., Tuttle, K. S., Zhou, S., Chittori, S., & Subramaniam, S. (2021). Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. *PLoS Biology*, *19*(4), 1–17. https://doi.org/10.1371/journal.pbio.3001237

Appendix 1.1: Derivatization of the rate equation for PPi production by a homodimer enzymes as a function of two substrates (GTP and 2´dGTP) assuming standard steady-state and preequilibrium.



Constants:

K_1

K_2

K_3

K_4

K_5

K_6

$Kcat_{cdiG}$

$Kcat_{cGdG}$

$Kcat_{cdidG}$

Arranging equations from the model:

1) $K_1 = \frac{[E][G]}{[E_G]}$ => $[E] = \frac{K_1[E_G]}{[G]}$

2) $K_2 = \frac{[E_G][G]}{[E_G^G]}$ => $[E_G] = \frac{K_2[E_G^G]}{[G]}$

3) $K_3 = \frac{[E][dG]}{[E_{dG}]}$ => $[E] = \frac{K_3[E_{dG}]}{[dG]}$

4) $K_4 = \frac{[E_{dG}][dG]}{[E_{dG}^{dG}]}$ => $[E_{dG}] = \frac{K_4[E_{dG}^{dG}]}{[dG]}$

5) $K_5 = \frac{[E_{dG}][G]}{[E_{dG}^G]}$ => $[E_{dG}] = \frac{K_5[E_{dG}^G]}{[G]}$

6) $K_6 = \frac{[E_G][dG]}{[E_{dG}^G]}$ => $[E_G] = \frac{K_6[E_{dG}^G]}{[dG]}$

7) *Replacing 2 in 1* => $[E] = \frac{\frac{K_1K_2[E_G^G]}{[G]}}{[G]}$ => $[E] = \frac{K_1K_2[E_G^G]}{[G]^2}$

8) *Replacing 4 in 3* => $[E] = \frac{\frac{K_3K_4[E_{dG}^{dG}]}{[dG]}}{[dG]}$ => $[E] = \frac{K_3K_4[E_{dG}^{dG}]}{[dG]^2}$

9) *Rearranging 1* => $[E_G] = \frac{[E][G]}{K_1}$

10) *Rearranging 3* => $[E_{dG}] = \frac{[E][dG]}{K_3}$

11) *Rearranging 5* => $[E_{dG}^G] = \frac{[E_{dG}][G]}{K_5}$

12) *Rearranging 6* => $[E_{dG}^G] = \frac{[E_G][dG]}{K_6}$

13) *Matching 12 and 11* => $\frac{[E_G][dG]}{K_6} = \frac{[E_{dG}][G]}{K_5}$

14) *Replacing 2 in 13* => $\frac{\frac{K_2[E_G^G][dG]}{[G]}}{K_6} = \frac{[E_{dG}][G]}{K_5}$ => $\frac{K_2K_5[E_G^G][dG]}{K_6[G^2]} = [E_{dG}]$

15) *Replacing 4 in 14* => $\frac{K_2K_5[E_G^G][dG]}{K_6[G]^2} = \frac{K_4[E_{dG}^{dG}]}{[dG]}$ => $\frac{K_2K_5[E_G^G][dG]^2}{K_4K_6[G]^2} = [E_{dG}^{dG}]$

16) $Replacing\ 2\ in\ 12 \quad => \quad \left[E_{dG}^{G}\right] = \frac{\frac{K_2\left[E_G^G\right][dG]}{[A]}}{K_6} \quad => \quad \left[E_{dG}^{G}\right] = \frac{K_2[E_G^G][dG]}{K_6[G]}$

17) $Replacing\ 9\ in\ 12 \quad => \quad \left[E_{dG}^{G}\right] = \frac{\frac{[E][G][dG]}{K_1}}{K_6} \quad => \quad [E] = \frac{K_1K_6[E_{dG}^G]}{[G][dG]}$

18) $Replacing\ 10\ in\ 11 \quad => \quad \left[E_{dG}^{G}\right] = \frac{\frac{[E][G][dG]}{K_3}}{K_5} \quad => \quad [E] = \frac{K_3K_5[E_{dG}^G]}{[G][dG]}$

19) $Rearranging\ 12 \quad => \quad [E_G] = \frac{K_6[E_{dG}^G]}{[dG]}$

20) $Replacing\ 2\ in\ 19 \quad => \quad \frac{K_2[E_G^G]}{[G]} = \frac{K_6[E_{dG}^G]}{[dG]} \quad => \quad [E_G^G] = \frac{K_6[E_{dG}^G][G]}{K_2[dG]}$

21) $Rearranging\ 11 \quad => \quad [E_{dG}] = \frac{K_5[E_{dG}^G]}{[G]}$

22) $Replacing\ 4\ in\ 21 \quad => \quad \frac{K_4[E_{dG}^{dG}]}{[dG]} = \frac{K_5[E_{dG}^G]}{[G]} \quad => \quad [E_{dG}^{dG}] = \frac{K_5[E_{dG}^G][dG]}{K_4[G]}$

23) $Replacing\ 4\ in\ 13 \quad => \quad \frac{[E_G][dG]}{K_6} = \frac{\frac{K_4\left[E_{dG}^{dG}\right][G]}{[dG]}}{K_5} \quad => \quad [E_G] = \frac{K_4K_6[E_{dG}^{dG}][G]}{K_5[dG]^2}$

24) $Replacing\ 2\ in\ 23 \quad => \quad \frac{K_2[E_G^G]}{[G]} = \frac{K_4K_6[E_{dG}^{dG}][G]}{K_5[dG]^2} \quad => \quad [E_G^G] = \frac{K_4K_6[E_{dG}^{dG}][G]^2}{K_2K_5[dG]^2}$

25) $Replacing\ 4\ in\ 11 \quad => \quad \left[E_{dG}^{G}\right] = \frac{\frac{K_4\left[E_{dG}^{dG}\right][G]}{[dG]}}{K_5} \quad => \quad \left[E_{dG}^{G}\right] = \frac{K_4[E_{dG}^{dG}][G]}{K_5[dG]}$

Total enzyme is:

26) $[E_t] = [E] + [E_G] + \left[E_G^G\right] + [E_{dG}] + \left[E_{dG}^{dG}\right] + [E_{dG}^G]$

Replacing Eq.26 in function of: $[E_G^G], [G], [dG], K_1, K_2, K_4, K_5, K_6$

$$[E_t] = (Eq.7) + (Eq.2) + [E_G^G] + (Eq.14) + (Eq.15) + (Eq.16)$$

$$[E_t] = \frac{K_1K_2[E_G^G]}{[G]^2} + \frac{K_2[E_G^G]}{[G]} + [E_G^G] + \frac{K_2K_5[E_G^G][dG]}{K_6[G]^2} + \frac{K_2K_5[E_G^G][dG]^2}{K_4K_6[G]^2} + \frac{K_2[E_G^G][dG]}{K_6[G]}$$

$$[E_t] = \frac{K_1K_2K_4K_6[E_G^G] + K_2K_4K_6[E_G^G][G] + K_4K_6[E_G^G][G]^2 + K_2K_4K_5[E_G^G][dG] + K_2K_5[E_G^G][dG]^2 + K_2K_4[E_G^G][G][dG]}{K_4K_6[G]^2}$$

$$[E_t] = \frac{[E_G^G](K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}{K_4K_6[G]^2}$$

Rearranging:

$$[E_G^G] = \frac{[E_t]K_4K_6[G]^2}{K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}$$

Considering:

$$Vo_{cdiG} = Kcat_{cdiG}[E_G^G]$$

$$Vmax_{cdiG} = Kcat_{cdiG}[E_t]$$

Multiplying by $Kcat_{cdiG}$:

27) $Vo_{cdiG} = \frac{Vmax_{cdiG}K_4K_6[G]^2}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$

We obtain the equation that describe initial velocity of formation of $cdiG$: Vo_{cdiG}

Now replacing Eq. 26 in function of: $[E_{dG}^{G}], [A], [B], K_1, K_2, K_4, K_5, K_6$

$$[E_t] = (Eq. 17) + (Eq. 19) + (Eq. 20) + (Eq. 21) + (Eq. 22) + [E_{dG}^{G}]$$

$$[E_t] = \frac{K_1K_6[E_{dG}^{G}]}{[G][dG]} + \frac{K_6[E_{dG}^{G}]}{[dG]} + \frac{K_6[E_{dG}^{G}][G]}{K_2[dG]} + \frac{K_5[E_{dG}^{G}]}{[G]} + \frac{K_5[E_{dG}^{G}][dG]}{K_4[G]} + [E_{dG}^{G}]$$

$$[E_t] = \frac{K_1K_2K_4K_6[E_{dG}^{G}] + K_2K_4K_6[E_{dG}^{G}][G] + K_4K_6[E_{dG}^{G}][G]^2 + K_2K_4K_5[E_{dG}^{G}][dG] + K_2K_5[E_{dG}^{G}][dG]^2 + K_2K_4[E_{dG}^{G}][G][dG]}{K_2K_4[G][dG]}$$

$$[E_t] = \frac{[E_{dG}^{G}](K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}{K_2K_4[G][dG]}$$

Rearranging:

$$[E_{dG}^{G}] = \frac{[E_t]K_2K_4[G][dG]}{K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}$$

Considering:

$$Vo_{cGdG} = Kcat_{cGdG}[E_{dG}^{G}]$$

$$Vmax_{cGdG} = Kcat_{cGdG}[E_t]$$

Multiplying by $Kcat_{cGdG}$:

28) $$Vo_{cGdG} = \frac{Vmax_{cdiG}K_2K_4[G][dG]}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$$

We obtain the equation that describe initial velocity of formation of $cGdG$: Vo_{cGdG}

Finally, replacing Eq. 26 in function of: $[E_{dG}^{dG}], [A], [B], K_2, K_3, K_4, K_5, K_6$

$$[E_t] = (Eq. 8) + (Eq. 23) + (Eq. 24) + (Eq. 4) + [E_{dG}^{dG}] + (Eq. 25)$$

$$[E_t] = \frac{K_3K_4[E_{dG}^{dG}]}{[dG]^2} + \frac{K_4K_6[E_{dG}^{dG}][G]}{K_5[dG]^2} + \frac{K_4K_6[E_{dG}^{dG}][G]^2}{K_2K_5[dG]^2} + \frac{K_4[E_{dG}^{dG}]}{[dG]} + [E_{dG}^{dG}] + \frac{K_4[E_{dG}^{dG}][G]}{K_5[dG]}$$

$$[E_t] = \frac{K_2K_3K_4K_5[E_{dG}^{dG}] + K_2K_4K_6[E_{dG}^{dG}][G] + K_4K_6[E_{dG}^{dG}][G]^2 + K_2K_4K_5[E_{dG}^{dG}][dG] + K_2K_5[E_{dG}^{dG}][dG]^2 + K_2K_4[E_{dG}^{dG}][G][dG]}{K_2K_5[dG]^2}$$

If $K_3K_5 = K_1K_6$:

$$[E_t] = \frac{K_1K_2K_4K_6[E_{dG}^{dG}] + K_2K_4K_6[E_{dG}^{dG}][G] + K_4K_6[E_{dG}^{dG}][G]^2 + K_2K_4K_5[E_{dG}^{dG}][dG] + K_2K_5[E_{dG}^{dG}][dG]^2 + K_2K_4[E_{dG}^{dG}][G][dG]}{K_2K_5[dG]^2}$$

$$[E_t] = \frac{[E_{dG}^{dG}](K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}{K_2K_5[dG]^2}$$

Rearranging:

$$[E_{dG}^{dG}] = \frac{[E_t]K_2K_5[dG]^2}{K_1K_2K_4K_6 + K_2K_4K_6[G] + K_4K_6[G]^2 + K_2K_4K_5[dG] + K_2K_5[dG]^2 + K_2K_4[G][dG]}$$

Considering:

$$Vo_{cdidG} = Kcat_{cdidG}[E_{dG}^{dG}]$$

$$Vmax_{cdidG} = Kcat_{cdidG}[E_t]$$

Multiplying by $Kcat_{cdidG}$:

29) $$Vo_{cdidG} = \frac{Vmax_{cdidG}K_2K_5[dG]^2}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$$

We obtain the equation that describe initial velocity of formation of $cdidG$: Vo_{cdidG}

Total initial velocity is the sum of the initial velocities for each product:

$$Vo_{total} = Vo_{cdiG} + Vo_{cGdG} + Vo_{cdidG}$$

Replacing:

$$Vo_{total} = (Eq.27) + (Eq.28) + (Eq.29)$$

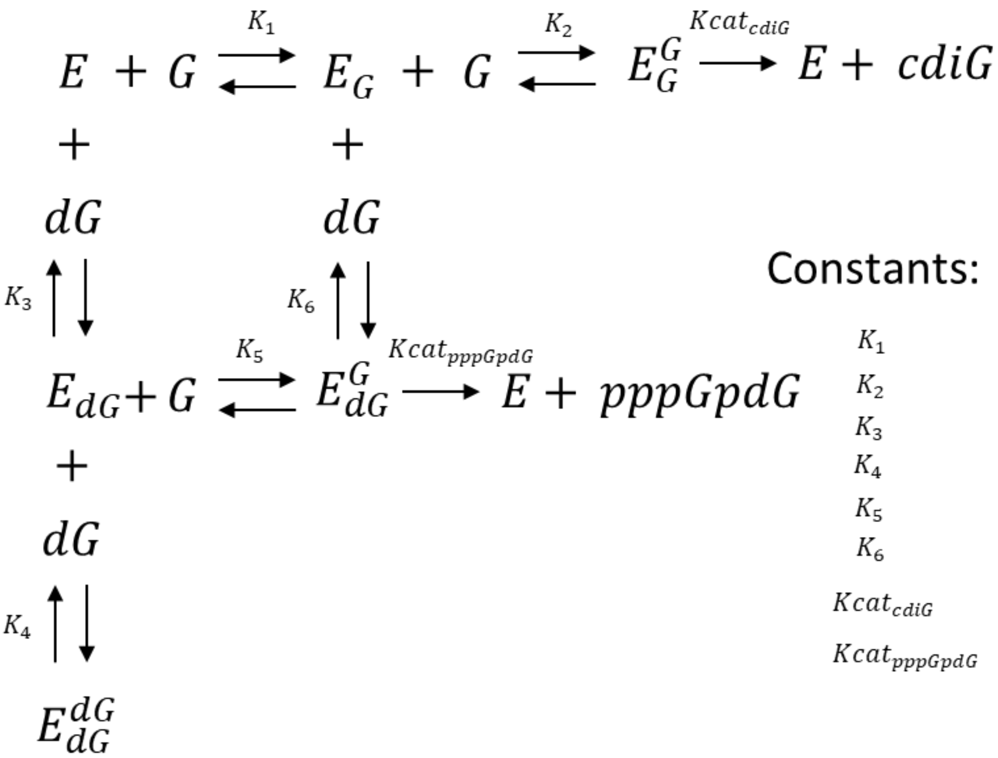
30) $$Vo_{total} = \frac{Vmax_{cdiG}K_4K_6[G]^2+Vmax_{cGdG}K_2K_4[G][dG]+Vmax_{cdidG}K_2K_5[dG]^2}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$$

Considering a model where: $Vmax_{cdidG} = 0$

31) $$Vo_{total} = \frac{Vmax_{cdiG}K_4K_6[G]^2+Vmax_{cGdG}K_2K_4[G][dG]}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$$

And, if $cGdG$ is not observed, instead we observe the linear product $pppGpdG$. Then, the reaction of formation of $cdiG$release two PPi and the reaction of formation of $pppGpdG$ release just one PPi. If we consider Vo_{total} in terms of $PPi\ x\ s^{-1}\ x\ XAC0610_{dimer}{}^{-1}$:

32) $$Vo_{total} = \frac{2\ Vmax_{cdiG}K_4K_6[G]^2+Vmax_{pppGpdG}K_2K_4[G][dG]}{K_1K_2K_4K_6+K_2K_4K_6[G]+K_4K_6[G]^2+K_2K_4K_5[dG]+K_2K_5[dG]^2+K_2K_4[G][dG]}$$



Assuming not-cooperativity $K_1 = K_2 = K_5 = K_G\ and\ K_3 = K_4 = K_6 = K_{dG}$:

33) $$Vo_{total} = \frac{2\ Vmax_{cdiG}K_{dG}{}^2[G]^2+Vmax_{pppGpdG}K_GK_{dG}[G][dG]}{K_G{}^2K_{dG}{}^2+K_GK_{dG}{}^2[G]+K_{dG}{}^2[G]^2+K_{dG}K_G{}^2[dG]+K_G{}^2[dG]^2+K_GK_{dG}[G][dG]}$$

Appendix 3.1: Santiago Justo Arevalo, Daniela Zapata Sifuentes, Cesar J. Huallpa, Gianfranco Landa Bianchi, Adriana Castillo Chavez, Romina Garavito-Salini Casas, Carmen Sofia Uribe Calampa, Guillermo Uceda-Campos, Roberto Pineda Chavarria. **Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures**. *Scientific reports* 11, 17755 (2021). https://doi.org/10.1038/s41598-021-97267-7

scientific reports

OPEN

Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures

Santiago Justo Arevalo[1,2]✉, Daniela Zapata Sifuentes[1], César J. Huallpa[3], Gianfranco Landa Bianchi[1], Adriana Castillo Chávez[1], Romina Garavito-Salini Casas[1], Carmen Sofia Uribe Calampa[1], Guillermo Uceda-Campos[2,4] & Roberto Pineda Chavarría[1]

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This disease has spread globally, causing more than 161.5 million cases and 3.3 million deaths to date. Surveillance and monitoring of new mutations in the virus' genome are crucial to our understanding of the adaptation of SARS-CoV-2. Moreover, how the temporal dynamics of these mutations is influenced by control measures and non-pharmaceutical interventions (NPIs) is poorly understood. Using 1,058,020 SARS-CoV-2 from sequenced COVID-19 cases from 98 countries (totaling 714 country-month combinations), we perform a normalization by COVID-19 cases to calculate the relative frequency of SARS-CoV-2 mutations and explore their dynamics over time. We found 115 mutations estimated to be present in more than 3% of global COVID-19 cases and determined three types of mutation dynamics: high-frequency, medium-frequency, and low-frequency. Classification of mutations based on temporal dynamics enable us to examine viral adaptation and evaluate the effects of implemented control measures in virus evolution during the pandemic. We showed that medium-frequency mutations are characterized by high prevalence in specific regions and/or in constant competition with other mutations in several regions. Finally, taking N501Y mutation as representative of high-frequency mutations, we showed that level of control measure stringency negatively correlates with the effective reproduction number of SARS-CoV-2 with high-frequency or not-high-frequency and both follows similar trends in different levels of stringency.

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a single-stranded positive RNA virus that infects humans. Since the first reported cases in December 2019, the disease has spread globally causing more than 161.5 million confirmed cases and 3.3 million deaths as of May 16th[1].

Since the emergence of COVID-19, significant genomic sequencing efforts have played a central role in furthering our understanding of the evolutionary dynamics of the virus. This has allowed the identification of mutations that appeared early in the pandemic (and that now seem to be fixed in the population[2–6]), as well as monitoring of the effectiveness of vaccines against variants coding for mutations in the spike[7–12]. Both underscore the importance of timely identification and surveillance of mutations with significant representation in the population, to efforts aimed at containing transmission of the virus.

[1]Facultad de Ciencias Biológicas, Universidad Ricardo Palma, Lima, Peru. [2]Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil. [3]Facultad de Ciencias, Universidad Nacional Agraria la Molina, Lima, Peru. [4]Facultad de Ciencias Biológicas, Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Peru. ✉email: santiago.justo@urp.edu.pe

The combination of virus spread by droplets through close contact[13,14], the large number of asymptomatic cases[15,16], the absence of effective pharmaceutical treatments at the beginning of the pandemic and the delays in production and distribution of vaccines[17], leave non-pharmaceutical interventions as the most effective measures to contain the spread of COVID-19 for a large fraction of the world's population.

Different studies have evaluated the relationship between non-pharmaceutical interventions (NPIs) and the decrease in the number of cases[18,19], the reproductive number[18,20,21], the case fatality rate[22], the contagion rate[23], and the number of SARS-CoV-2 importations[24,25]. By contrast, the effect of NPIs on specific mutations has been less well-studied. Pachetti et al.[22] analyzed how lockdown policies might have influenced the dynamics of some SARS-CoV-2 mutations; however, results are primarily qualitative and little quantitative description of the reported effect is provided. Muller et al.[26] use phylogenetic methods to estimate the importance of SARS-CoV-2 introductions on increasing the relative frequency of the D614G mutation, implicitly showing that international movement can affect the relative frequency of mutations.

Here, using 1,058,020 genomes from sequenced COVID-19 cases, we analyze the temporal dynamics of SARS-CoV-2 mutations estimated to be present in more than 3% of global COVID-19 cases. We then investigate whether mutations are region-specific and if there is a correlation between level of lockdown policies and the effective reproduction number of specific mutations.

Results and discussions

115 mutations overpass presence on 3% of global COVID-19 cases and most of them are non-synonymous.

We performed a by case normalization of the frequencies of the mutations from 1,058,020 genomes all around the world. The relative frequency of cases where a mutation is present was named Normalized Relative Frequency of a genomic position: NRFp. The NRFp of each mutation was calculated from genomes and the number of cases of 714 country-month combinations, including 98 countries from January 2020 to April 2021.

This normalization allowed us to identify mutations that have not been reported in other global studies, such as that of Castonguay et al.[27]. This is because in many countries the number of sequenced genomes is low and certain mutations could go unnoticed. Thus, we identified 115 mutations with NRFp > 0.03 (Fig. S1); this means that those mutations are estimated to be present in more than 3% of the COVID-19 cases globally. Considering that the sum of the reported cases from the 714 country-month combinations analyzed was 120,008,410 cases, an NRFp of more than 0.03 means that those mutations were present in more than 3,600,252 global COVID-19 cases.

Table S1 summarizes the features of these 115 mutations. Based on those 115 mutations, we calculated a dN/dS ratio of 4.1 that could imply positive selection occurring in the SARS-CoV-2 genome. Additionally, S and N proteins did not show synonymous mutations and presents ~ 74% of the total non-synonymous mutations suggesting that positive selection is predominantly in those two ORFs.

Mutations show three types of temporal dynamics.

The dynamics of the 115 mutations were analyzed through calculating the NRFp in each month from January 2020 to April 2021 (Fig. 1). We assigned type of temporal dynamics to the mutations according to the NRFp in different months and the change of NRFp between months. Thus, three types of temporal dynamics were observed: (i) high-frequency mutations (HF) that never show negative NRFp changes greater than 1%, and increased rapidly in NRFp since their appearance (Fig. 1a), (ii) medium-frequency mutations (MF) that alternates between negative and positive NRFp changes and presents at least one month with NRFp greater than 15% (Fig. 1b), and (iii) low-frequency mutations (LF) that also have an alternation between negative and positive NRFp changes but at a NRFp ever below 15% (Fig. 1c).

HF mutations are characterized by a rapid increase in global frequency following their appearance (Fig. 1a). This could be due to positive selection without competition and/or by other effects related to population dynamics such as control measures implemented by countries aimed at controlling transmission. Mutations in this category appeared in two well-defined stages of the pandemic. The first group is composed of four mutations that now appear to be globally fixed. They emerged at the beginning of the pandemic in January 2020, reaching more than 0.75 NRFp in April 2020 (Fig. 1a, Group 1). The second group rapidly increased in frequency in December 2020, and have continued to increase since then (Fig. 1a, Group 2).

Some HF mutations identified here have been widely reported[28] due to their presence in variants of concern. The first and second groups contained Spike mutations well known due to their possible implications in transmissibility, (e.g. D614G in the first group[29,30] (Fig. S2b)), and vaccine efficacy (e.g. Δ69–70, N501Y, and E484K, all present in the second group[31,32] (Fig. S2b,e)). In the future, analysis of the dynamics of other mutations in this way could help facilitate rapid identification of other mutations of concern.

By contrast, some of the MF and LF mutations that we observed have not been less previously reported to a significant degree, with descriptions either limited to specific countries or regions[33–35], or not reported at all, (e.g. K997Q on nsp3 and S202C on N protein). However, those mutations are present in several months throughout the pandemic and we did not observe evidence of the extinction of any of these mutations (relative frequency of 0 or near to 0 in two or more consecutive months) (Fig. 1b,c).

One possibility for the existence of MF and LF mutations is that some benefits may be conferred to SARS-CoV-2 but competition with other variants prevents rapid increases in their frequency increase across the population. Such dynamics have been observed in evolution experiments for other organisms[36,37]. Furthermore, the coexistence of different lineages of the same organism in the context of frequency-dependent interactions has been reported in yeast[38] and bacteria[39,40], and have highlighted that this can be beneficial for the organism. In the case of virus, epitope diversity and host-specific adaptation can be beneficial for the viral population[41].

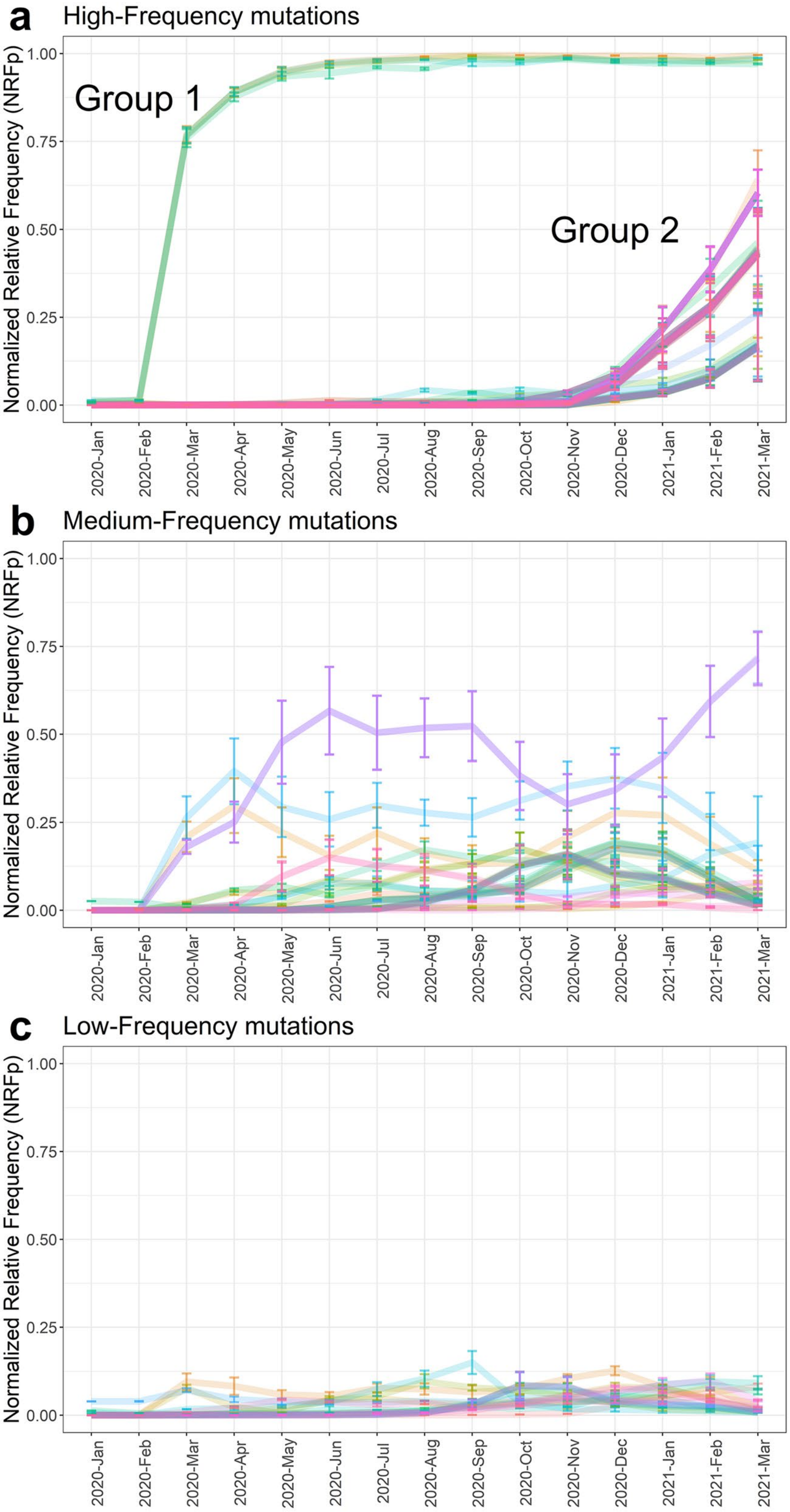


Figure 1. Three different temporal dynamics of SARS-CoV-2 mutations. Normalized by cases Relative Frequency (NRFp) of the mutations by month. **(a)** high-frequency mutations (HF) never show negative NRFp changes greater than 1%, and increased rapidly since their appearance. **(b)** medium-frequency mutations (MF) alternates between negative and positive NRFp changes and presents at least one month with NRFp greater than 15% **(c)** low-frequency mutations (LF) that also have an alternation between negative and positive NRFp changes but at a NRFp ever below 15%. Error bars represent inter-region variation as weighted variance.

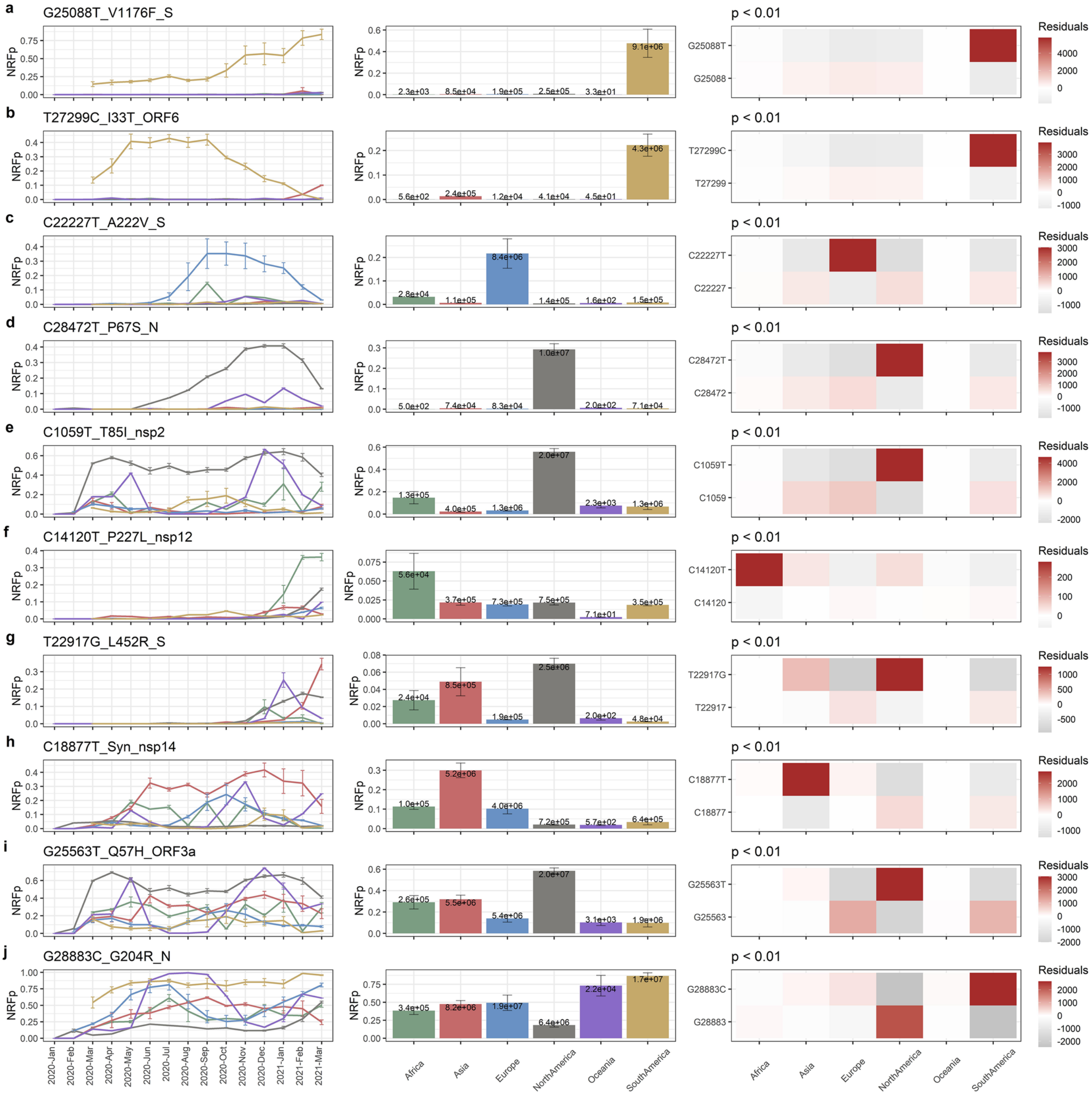


Figure 2. MF mutations are region-specific or have mid-frequencies in several regions. (Left-column) Normalized by cases Relative Frequency (NRFp) of the mutations by month separated by regions (green = Africa, red = Asia, blue = Europe, grey = North America, purple = Oceania, yellow = South America). (Middle-column) Total NRFp by region of the analyzed medium-frequency mutations. Numbers in each bar represents the estimated total number of cases of the particular mutation in that region. (Right-column) Chi-square p-value and Pearson residuals analysis of medium-frequency mutations. Upper line corresponds to the mutant state and the bottom line to the not-mutant state. Grey and red boxes mean negative or positive association with the state, respectively. Intensity of the colors means higher residuals that means greater contribution. **(a–e)** Region-specific medium-frequency mutations and **(f–j)** not-region-specific medium-frequency mutations.

Some of the MF mutations are region-specific while other have medium frequencies in various regions. In our previous work[42], we observed that the mutation T85I in nsp2 has a higher frequency in North America than in other continents. Here, we show that this MF mutation maintains this tendency, persisting since its appearance at a global NRFp of ~ 0.2 (Fig. S3a). Interestingly, and in contrast to HF mutations (that are typically similarly frequent across several analyzed regions, with an exception being a group of recent mutations that are more frequent in South America (Figs. S4, S5)), most of the MF mutations (18 of 29) analyzed here are most frequent in a specific region (Fig. 2).

To explore whether MF mutations showed a region-specific pattern, we analyzed the dynamics of ten subtypes of MF dynamics in six different regions (Africa, Asia, Europe, North America, Oceania, and South America) (Fig. S6). Our results show that five subtypes had a NRFp greater than 0.3 for at least three consecutive months in only one region (Fig. 2a–e, left column). Relatedly, mutations belonging to these subtypes had a higher relative number of cumulative cases (NRFp) in a specific region, compared to other regions (Fig. 2a–e, middle column).

Then, we examined whether the proportions of estimated COVID-19 cases caused by MF mutations were different between regions. Chi-square p-values showed that in all the MF subtypes at least one region have different proportions (Fig. 2, right column). Pearson residuals analysis showed which of the regions have larger or smaller mutant proportion than expected (meaning positive or negative association, red and grey squares, respectively) and which region has a greater degree of association (color intensity). The five subtypes that showed region-specific patterns also showed that just one region is positively associated and that it has the highest degree of association to that specific mutation (Fig. 2a–e, right column). By contrast, other five subtypes showed positive association to more than one region with a variety of degrees of association (Fig. 2f–j, right column).

We further analyze the five subtypes that showed region-specific pattern (Fig. 2a–e). Country analysis of the relative frequencies (Figs. S7, S8) and the cumulative number of cases (Figs. S9, S10) showed that those mutations are found in more than one country of the region. Some of them follows a similar pattern of frequency changes in two or more countries within the region (S7b, S7d and S8), whereas others have a particular pattern of frequency change in one particular country (S7a and S7c). Analysis of the cumulative number of cases by country showed that, although several countries present COVID-19 cases of the particular mutations, in most cases few countries contributes to most cases (S9a, Brazil; S9b, Argentina, Brazil, Chile; S9c, USA; S9d, Canada, Mexico, USA; S10, Italy, Spain, UK).

A decline of the frequency can be seen for some MF mutations in the last months (Fig. 2b–d), this can be explained because new mutations leave out of competition those mutations, or due to a delay between the collection date and the submission date of genomic samples. Using genomic data from August 10th 2021, we re-analyzed three mutations that clearly showed this decline (Fig. 2b (I33T), c (A222V), and d (P67S)). We found very similar patterns in the countries analyzed (Fig. S11, S12), therefore, leave out of competition by other mutations is a more plausible scenario.

LF mutations followed similar patterns to those observed for MF mutations (Figs. S13, S14). Thus, the MF and LF dynamics seems to be due to: (i) high prevalence of mutations in specific regions, (ii) globally dispersed beneficial mutations in constant competition with other variants, or (iii) a combination of these two effects.

SARS-CoV-2 carrying HF_{N501Y} mutation follows similar trends than SARS-CoV-2 without HF_{N501Y} in different levels of control measures. The rapid increase in global frequency of HF mutations and the observation that those mutations appeared at two very defined stages of the pandemic (Fig. 1a) lead us to hypothesize that, at least part of this abrupt increase is due to the fact that limited or minimal levels of control measures and NPIs may permit that HF mutations to spread even faster than not-HF mutations that when stronger control measures and stronger NPIs are present. An alternative hypothesis could be that strict control measures give a large competitive advantage to more transmissible variants (HF mutations), enabling them to persist and continuing to transmit, whilst their less transmissible counterparts (not-HF mutations) die out.

To test these hypotheses, we analyzed whether different degree of control measures could affect differently to SARS-CoV-2 genomes bearing the HF mutation N501Y (HF_{N501Y}) or not bearing the HF mutation N501Y (not-HF_{N501Y}) in nine countries that have more than 15 sequenced genomes per week during March 2020 to April 2021. We selected this mutation because it is present in three variants of concern (B.1.1.7, B.1.35, and P.1)[43] and is a good example of the behavior of HF mutations (Fig. 1a, Supplementary Fig. S2a). Additionally, and in contrast to HF mutations belonging to group 2, mutations in the first group of HF mutations (Fig. 1a, group 1) may have been aided by founder effects in the early stages of the pandemic. For this reason, we did not analyze them in this part of our study.

First, we estimated the effective reproduction number (Rt) of HF_{N501Y} or not-HF_{N501Y} (Fig. 3a) and measure the correlation with the level of stringency (Fig. 3b). The level of stringency is a measure of the level of control policies based on nine response indicators including school closing, workplace closing, cancel public events, restrictions on gathering size, close public transport, stay-at-home requirements, restrictions on internal movement, restriction on international travel and public information campaigns[44].

We found significant negative correlation between the Rt after 14 days that the level of stringency was implemented and the level of stringency in eight of the nine countries analyzed (Fig. 3b). In all these eight countries linear regression model explained at least 23% of the variance in the Rt of HF_{N501Y} (Fig. 3b), and the effect size measured by the R-value of spearman correlation showed in the worst case a value of 0.48, with all the others R-value between 0.5 and 0.81 (Fig. 3b). In the case of India, the Rt of HF_{N501Y} showed a positive correlation with level of stringency. It is known that efforts in molecular testing in India have changed during the pandemic[45] Time-varying differences in the intensity and capacity of molecular testing can produce significant biases in the estimation of Rt. Overall however, our results show a significant negative correlation between degree of control measure stringency and Rt in eight of the nine countries analyzed.

We also found that, independently of the level of stringency imposed, the Rt of HF_{N501Y} was significantly higher than not-HF_{N501Y}, potentially explaining why HF_{N501Y} increase its frequency faster than not-HF_{N501Y} since its appearance in the nine countries considered here (Fig. 4a). Interestingly, when we analyzed the Rt of SARS-CoV-2 genomes bearing an MF mutation (MF_{R203K}) and compare it with the Rt of genomes without the MF mutation (not-MF_{R203K}) we observed that in some stages of the pandemic the Rt of MF_{R203K} is higher than not-MF_{R203K} but in other cases the opposite was observed (Fig. S15). This explains why this mutation did not increases its frequency steadily and can be an evidence of constant competition between MF_{R203K} and not-MF_{R203K}.

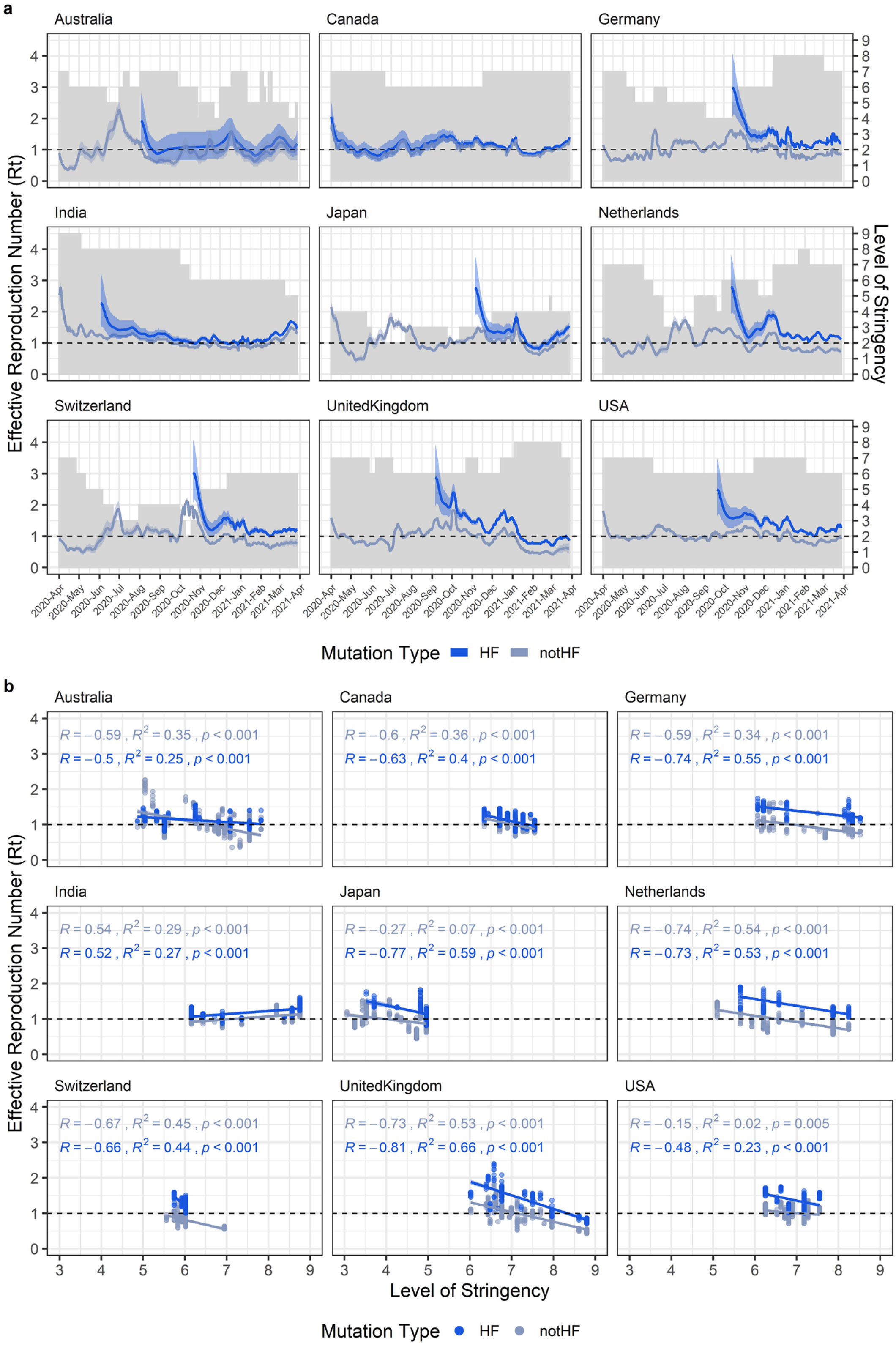


Figure 3. Effective reproduction number (Rt) of HF_{N501Y} and not-HFN_{501Y} are correlated with level of stringency. **(a)** Each panel shows the estimated effective reproduction number of SARS-CoV-2 bearing (blue) or not (grey) the HF mutation N501Y (HF or notHF, respectively) in different countries. Grey bars are showing the level of stringency. Shades show a 97.5% confidence interval in the estimation of Rt. **(b)** Correlation of Rt after 14 days of the implementation of the level of stringency with the level of stringency. Each panel shows the independent analysis of different countries. Spearman correlation values (R), R-square of the linear regression model (R^2), and p-value of the correlation is showed in the left-up of each panel in this order. Colors represent the same as in **(a)**.

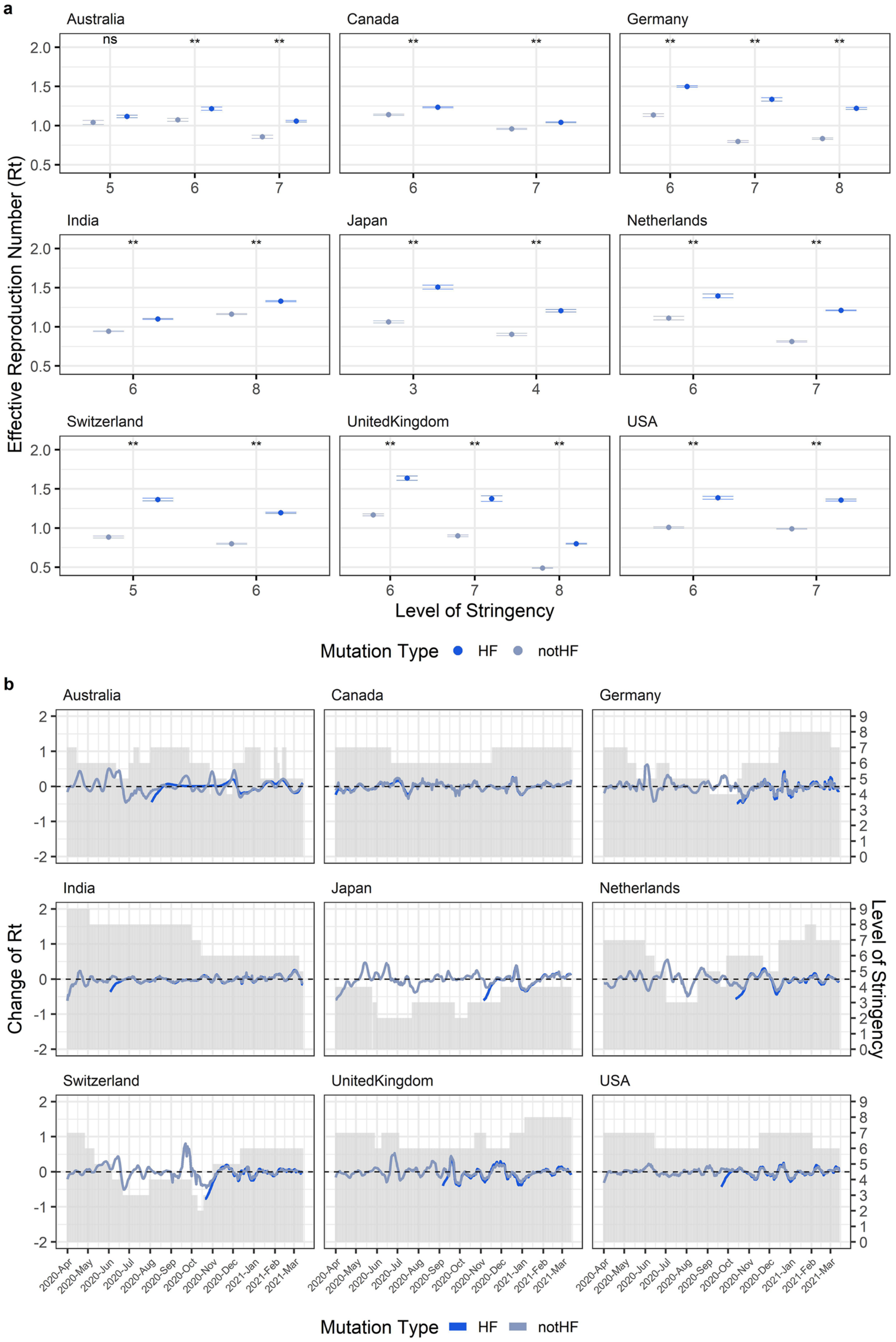


Figure 4. Effective reproduction number (Rt) of HF_{N501Y} is higher than not-HF_{N501Y} but similarly affected by the level of stringency. **(a)** Statistical comparison between the bootstrap distribution of the Rt of SARS-CoV-2 bearing (blue) or not (grey) the HF mutation N501Y (HF or notHF, respectively) in different levels of stringency. Points represent the mean and the lines represent the 25 and 75 percentiles of the bootstrap distribution. **Means p-value lesser than 0.05 and ns means p-value higher than 0.05. **(b)** Plot of the change of Rt in the time. Change of Rt was calculated as the Rt 14 days after the day of interest subtracted to the Rt mean between the day of interest and 13 days after that day. Grey bars are showing the level of stringency. Colors represent the same as in **(a)**.

Finally, to explore whether different stringency levels differentially affect the dynamics and transmission of HF_{N501Y} and not-HF_{N501Y}, we calculated the change in Rt during several months where different levels of stringency were implemented (Fig. 4b). The patterns of the change of Rt between HF_{N501Y} and not-HF_{N501Y} were almost identical (Fig. 4b) and R values of spearman correlation of HF_{N501Y} and not-HF_{N501Y} with Rt were similar in most cases (Fig. 3b), indicating that both could be similarly affected by the changes in stringency levels.

Taken together, although HF_{N501Y} presented higher Rt in lower levels of stringency indicating that HF_{N501Y} spread was likely helped by mild lockdown policies in some stages of the pandemic, this effect was also observed in not-HF_{N501Y}. In conclusion, the results of this section showed control measures and their associated stringency probably affecting HF_{N501Y} and not-HF_{N501Y} in a similar fashion; thus, our two initial hypotheses are not supported by these results. Instead, the rapid increase of frequency of HF_{N501Y} is justified primarily by its generally increased transmissibility (i.e. a higher Rt which is always greater than the Rt of not-HF_{N501Y}), rather than the implementation of specific control measures.

Limitations of the study.

Stringency level is calculated from set of policies applied in each country that do not necessarily operate or function the same in different countries due to, for instance, variations in sociocultural and economic factors. Thus, comparisons at country level have variation that limit the reliability and interpretability of the results presented here, especially when compared with other countries. Moreover, different combinations of policies can generate the same level of stringency—the fact that several policies were applied together to generate a stringency index precludes efforts to evaluate the effect of a specific policy on the effective reproduction number of SARS-CoV-2.

After control measures are implemented (reflected as an increase to the stringency index) Rt changes from a higher value to a lower value. This process generates a time-window of intermediate Rt before the Rt reach a plateau that indicates how much the policy lowered the Rt. These intermediate values of Rt introduce a bias in the correlation between Rt and the level of stringency. Furthermore, if a country changes the stringency level in time-windows less than those necessary for the Rt to stabilize, the estimations of correlation get more complicated.

Our correlation analysis showed that in seven of the nine countries analyzed lower levels of stringency are correlated with higher Rt values. This could be an evidence of a possible effect of lockdown policies in the Rt. However, causal inference model is known to be a more accurate approach to test causality.

Although the methodology of normalization by cases alleviates the differences in the number of genomes sequenced by country, confidence in the calculation of relative frequencies of mutations is still low in regions with a low number of genomes sequenced. For example, a mutation with 0.5 relative frequency that comes from a sample of 15 genomes will have a confidence interval between 0.25 and 0.75; on the other hand, a sample of 150 genomes will generate a confidence interval between 0.58 and 0.42. Also, the number of cases is still subjected to bias due to for instance, the difference in the number of tests that each country performs, as occurs in India.

Conclusions

Normalization by cases of the frequency of mutations is an important tool for global analyses in a pandemic where not all the countries possess the same capacity to sequence SARS-CoV-2 genomes. This process partially mitigates differences in available genomes, but does not eliminate this problem. Worldwide efforts to help countries with fewer sequencing resources would improve our understanding of the adaptation and evolution process of SARS-CoV-2.

Three types of dynamics of mutations are described here and named “high-frequency” (HF), “medium-frequency” (MF), and “low-frequency” (LF). The three types are represented in all the months analyzed, and found in non-structural and structural proteins, and synonymous and non-synonymous mutations. Differences in the dynamics could be due to different forces acting on each of these types of mutations and the implications of all of them need to be studied to better understand the adaptation process of SARS-CoV-2.

Medium and low-frequency mutations maintain roughly constants global frequency due to their higher prevalence on specific regions and/or because they are in constant competition with other mutations in several regions. We showed some mutations with a high degree of region-specificity and others that presented mid-frequencies in several regions. Higher prevalence in specific regions may be due to specific-host characteristics. Constant competition in several regions may be due to the fact that they are beneficial mutation in the presence of other mutations with a similar degree of benefit. Some mutation can be leave out of competition when others beneficial mutations appear. Our analysis, also shows evidence that some MF mutations have a reduced relative frequency after several months of high frequencies in a specific region.

In this pandemic, human behavior has strongly affected the adaptive process of the SARS-CoV-2 through continuous implementations and changes to implemented control measures. Our analysis presents evidence that the high-frequency mutation N501Y is more transmissible (showed for its greater effective reproduction number) than not-N501Y, but also that control measures do not significantly favor the growth of any one in particular. Instead, we observe that policies have a similar impact on both.

Methods

Normalized by cases relative frequency of mutations on the SARS-CoV-2 genome.

To perform mutation frequency analysis considering the number of cases in each country we followed similar steps as described in Justo et al.[42], with some modifications: we first downloaded 1,221,746 genomes from the GISAID database (as of April 24th, 2021). Sequences with less than 29,000 nt were removed and the resulting sequences were aligned against the reference SARS-CoV-2 genome (EPI_ISL_402125) from nt 203 to nt 29,674 using ViralMSA.py[46,47]. From this alignment, we removed sequences with more than 290 Ns, more than 0.05% unique mutations, and/or more than 2% gaps. After those filters, we had 1,058,020 genomes. Subalignments were generated by grouping sequences by country and month. Subalignments with less than 15 sequences were not considered in the analysis. Nucleotide relative frequencies of each genomic position on each of 714 subalignments each corresponding to a different country-month combination (including 98 countries) were calculated. Normalized relative frequencies (NRFp) were calculated as the weighted mean of the relative frequencies in each subalignment with the number of cases as the weight. The number of cases for each month and country was obtained from the European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide). The NRFp is an estimation of the percentage of global COVID-19 cases where a particular mutation is present. The same procedure was done to obtain the NRFp of the mutations by months or by regions. Data manipulation was done using R and python scripts.

Analysis of region-specific mutations. The frequencies by country-months of each mutation were obtained from the previous calculation. Then, the Normalized relative frequencies (NRFp) by region (Africa, Asia, Europe, North America, Oceania, South America) were calculated as the weighted mean of the relative frequencies of each country-month belonging to a specific region using the number of cases as the weight. Number of cases with a particular mutation in each country was estimated by multiplying the relative frequency of the mutation with the number of cases in a specific country-month. Then, we added the cases belonging to a specific region and chi-square analyses were done using R software[48].

Estimation of effective reproduction number of SARS-CoV-2 mutations. We select nine countries (Australia, Canada, Germany, India, Japan, Netherlands, Switzerland, United Kingdom, USA) with at least 15 sequenced genomes by week from March 2020 to March 2021. Raw number of cases by days were obtained from[49] and used to estimate the number of cases by day for a specific mutation. In the case of MF mutation R203K, R203, and N501, we multiply the relative frequencies of the genomes with the state of interest (R203K, R203 or N501) in a determined week by the number of cases in the day. For instance, if 1 week presented 30% of genomes with the mutation R203K, and the number of cases on Monday of that week was 100. Thus, the estimated number of cases with this mutation in that day was 30. In the case of the HF mutation N501Y we first calculated the relative frequencies of that mutation in each week and then adjusted the relative frequencies to a logistic regression model using R software[50]. The number of cases estimated for the MF and HF mutations by day were used to estimate the effective reproduction number using EpiFilter[51].

Correlation analysis between stringency levels and effective reproduction number. The stringency index by country by day was obtained from[49]. Analysis of Spearman correlations and linear regression models of the effective reproduction number 14 days after the level of stringency was implemented with stringency index in each country by each state (mutant or not mutant) was done using R[48] and the packages ggplot2[52] and ggpubr[50].

Statistical differences between effective reproduction number of SARS-CoV-2 mutations in different levels of stringency. To determine if SARS-CoV-2 with HF_{N501Y} and not-HF_{N501Y} mutations presented statistical differences in Rt in different levels of stringency, we categorize the stringency index in ten levels: 0–10 = 0, 11–20 = 1, 21–30 = 2, 31–40 = 3, 41–50 = 4, 51–60 = 5, 61–70 = 6, 71–80 = 7, 81–90 = 8, and 91–100 = 9. We estimated the distribution of the effective reproduction number 14 days after the level of stringency was implemented in each level of stringency by bootstrap using 1000 replicates. Level of stringency with at least 10 Rt points were considered in the bootstrap analysis. We also used bootstrap methods to estimate the distribution of the difference of the Rt assuming that both Rt (HF_{N501Y} and not-HF_{N501Y}) comes from the same distribution and calculate the p-value of the observed difference.

Calculation of change in time of the effective reproduction number of SARS-CoV-2 mutations. Change of Rt was calculated by subtracting the value of Rt 14 days after the day of interest with the mean of the Rt from the day of interest to 13 days after the day of interest.

Data availability

Publicly available datasets were analyzed in this study. This data can be found at: gisaid.org. All the code used to perform the analysis of this manuscript is publicly available in: https://github.com/sanjusare/Justo_et_al_2021_SR.

Received: 30 November 2020; Accepted: 24 August 2021
Published online: 07 September 2021

References

1. World Health Organization (2021). https://covid19.who.int/ (Accessed 16 May 2021).
2. Korber, B. *et al.* Tracking changes in SARS-Cov-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827 (2020).
3. Biswas, S. K. & Mudi, S. R. Spike protein d614g and RDRP p323l: The SARS-CoV-2 mutations associated with severity of COVID-19. *Genomics Inform.* **18**(4), 1–7. https://doi.org/10.5808/GI.2020.18.4.e44 (2020).
4. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *BioRxiv.* https://doi.org/10.1101/2020.04.29.069054 (2020).
5. Yurkovetskiy, L. *et al.* Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**(3), 739–751. https://doi.org/10.1016/j.cell.2020.09.032 (2020).

6. Callaway, E. Making sense of coronavirus mutations. *Nature* **585**, 174–177 (2020).
7. Khan, A. *et al.* Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. *J. Cell. Physiol.* https://doi.org/10.1002/jcp.30367 (2021).
8. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv.* https://doi.org/10.1101/2020.12.21.20248640v1 (2020).
9. Jangra, S. *et al.* SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe.* https://doi.org/10.1016/S2666-5247(21)00068-9 (2021).
10. Leung, K., Shum, M. H. H., Leung, G. M., Lam, T. T. Y. & Wu, J. T. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Eurosurveillance.* https://doi.org/10.2807/1560-7917.ES.2020.26.1.2002106 (2021).
11. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 transmission. *BioRxiv.* https://doi.org/10.1101/2021.03.08.434499 (2021).
12. Kemp, S. *et al.* Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/70. *BioRxiv.* https://doi.org/10.1101/2020.12.14.422555v6 (2021).
13. Santarpia, J. *et al.* Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care. *Sci. Rep.* **10**, 12732 (2020).
14. Leung, N. *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Med.* **26**, 676–680 (2020).
15. Sayampanathan, A. *et al.* Infectivity of asymptomatic versus symptomatic COVID-19. *The Lancet* **397**(10269), 93–94 (2021).
16. Alene, M. *et al.* Magnitud of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLoS ONE* **16**(3), e0249090. https://doi.org/10.1371/journal.pone.0249090 (2021).
17. Kim, J., Marks, F. & Clemens, J. Looking beyond COVID-19 vaccine phase 3 trials. *Nat. Med.* **27**, 205–211 (2021).
18. Hyafil, A. & Moriña, D. Analysis of the impact of lockdown on the reproduction number of the SARS-CoV-2 in Spain. *Gac. Sanit.* https://doi.org/10.1016/j.gaceta.2020.05.003 (2020).
19. Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**(7820), 262–267 (2020).
20. Agrawal, M., Kanitkar, M. & Vidyasagar, M. Modelling the spread of SARS-CoV-2 pandemic-Impact of lockdowns & interventions. *Indian J. Med. Res.* **153**, 175–181 (2021).
21. Liu, Y., Morgenstern, C., Kelly, J., Lowe, R. & Jit, M. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Med.* **19**(1), 1–12 (2021).
22. Pachetti, M. *et al.* Impact of lockdown on Covid-19 case fatality rate and viral mutations spread in 7 countries in Europe and North America. *J. Transl. Med.* **18**, 338 (2020).
23. Larrosa, J. M. SARS-CoV-2 in Argentina: Lockdown, mobility, and contagion. *J. Med. Virol.* **93**(4), 2252–2261 (2021).
24. Adekunle, A., Meehan, M., Rojas-Alvarez, D., Trauer, J. & McBryde, E. Delaying the COVID-19 epidemic in Australia: Evaluation of the effectiveness of international travel bans. *Aust. N. Z. J. Public Health* **44**(4), 257–259. https://doi.org/10.1111/1753-6405.13016 (2020).
25. Wells, C. R. *et al.* Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proc. Natl. Acad. Sci.* **117**(13), 7504–7509 (2020).
26. Muller, N. F. *et al.* Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci. Transl. Med.* https://doi.org/10.1126/scitranslmed.abf0202 (2021).
27. Castonguay, N., Zhang, W. & Langlois, M. Meta-analysis and structural dynamics of the emergence of genetic variants of SARS-CoV-2. *MedRxiv.* https://doi.org/10.1101/2021.03.06.21252994v2 (2021).
28. Plante, J. *et al.* The variant gambit: COVID-19's next move. *Cell Host Microbe* **29**, 508 (2021).
29. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122–127 (2021).
30. Hou, Y. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).
31. Supasa, P. *et al.* Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* **184**, 2201–2211 (2021).
32. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939–2954 (2021).
33. Hodcroft, E. E. *et al.* Emergence of multiple lineages of SARS-CoV-2 spike protein variants affecting amino acid position 677. *MedRxiv.* https://doi.org/10.1101/2021.02.12.21251658v3 (2020).
34. Nagy, Á., Pongor, S. & Győrffy, B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. *Int. J. Antimicrob. Agents* **57**, 106272. https://doi.org/10.1016/j.ijantimicag.2020.106272 (2021).
35. Zrelovs, N. *et al.* First report on the latvian SARS-CoV-2 isolate genetic diversity. *Front. Med.* **8**, 626000. https://doi.org/10.3389/fmed.2021.626000 (2021).
36. Good, B., McDonald, M., Barrick, J., Lenski, R. & Desai, M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
37. Luksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
38. Frenkel, E. *et al.* Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. *PNAS* **112**(36), 11306–11311 (2015).
39. Maddamsetti, R., Lenski, R. & Barrick, J. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* **200**, 619–631 (2015).
40. Rozen, D. & Lenski, R. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *Am. Nat.* **155**(1), 24–35 (2000).
41. Parameswaran, P. *et al.* Intrahost selection pressures drive rapid dengue virus microevolution in acute human infections. *Cell Host Microbe* **22**, 400–410 (2017).
42. Justo, S. *et al.* Global geographic and temporal analysis of SARS-CoV-2 haplotypes normalized by COVID-19 cases during the pandemic. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2021.612432 (2021).
43. Konings, F. *et al.* SARS-CoV-2 Variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823 (2021).
44. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).
45. Unnikrishnan, J., Mangalathu, S. & Kutty, R. Estimating under-reporting of COVID-19 cases in Indian states: An approach using a delay-adjusted case fatality ratio. *BMJ Open* **11**, e042584 (2021).
46. Moshiri, N. ViralMSA: Massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* **37**(5), 714–716 (2020).
47. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
48. Core Team (2021). *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021). http://www.R-project.org/ (Accessed 16 May 2021).
49. Ritchie, H., *et al. Coronavirus Pandemic (COVID-19)* (2020). https://ourworldindata.org/coronavirus (Accessed 16 May 2021).
50. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots* (2020). https://CRAN.R-project.org/package=ggpubr (Accessed 16 May 2021).

51. Parag, K. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *MedRxiv.* https://doi.org/10.1101/2020.09.14.20194589v3 (2021).
52. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016) (Accessed 25 July 2021).

Acknowledgements

We are very grateful to the GISAID Initiative and all its data contributors, i.e., the authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. We thank Prof. Jose Luis Mena (Faculty of Biological Sciences—Universidad Ricardo Palma) and PhD(c). Carlos Prete (Escola Politecnica—University of Sao Paulo) for their valuable help in the epidemiological and statistical analysis. We thank Prof. Shaker Chuck Farah (Institute of Chemistry—University of Sao Paulo) and PhD(c). Charlie Whittaker (Faculty of Medicine—Imperial College London) for English writing corrections and helpful comments. To the Ricardo Palma University High-Performance Computational Cluster (URPHPC) managers Gustavo Adolfo Abarca Valdiviezo and Roxana Paola Mier Hermoza at the Ricardo Palma Informatic Department (OFICIC) for their contribution in programs and remote use configuration of URPHPC. Also, to one of the managers of the High-Performance Computer from the Instituto de Investigaciones de la Amazonía Peruana, Rodolfo Cardena Vigo for its assistance in configurations and program installations.

Author contributions

S.J.A. designed the study. S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C., R.G.-S.C., C.S.U.C. and R.P.C. analyzed the data. S.J.A., C.S.U.C. and G.U.-C. wrote python and R scripts. The manuscript was written by S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C., R.G.-S.C. and C.S.U.C. All authors discussed the methodologies, results, and read and approved the manuscript.

Funding

We thank to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) graduate scolarship (to SA; 2015/13318-4) and Universidad Ricardo Palma (URP) for APC financing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97267-7.

Correspondence and requests for materials should be addressed to S.J.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 3.2: Santiago Justo Arevalo, Daniela Zapata Sifuentes, Cesar J. Huallpa, Gianfranco Landa Bianchi, Adriana Castillo Chavez, Romina Garavito-Salini Casas, Guillermo Uceda-Campos, Roberto Pineda Chavarria. **Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes normalized by COVID-19 cases during the pandemic**. *Frontiers in Microbiology* 12, 612432 (2021). https://doi.org/10.3389/fmicb.2021.612432

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 February 2021
doi: 10.3389/fmicb.2021.612432

Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes Normalized by COVID-19 Cases During the Pandemic

Santiago Justo Arevalo[1,2], Daniela Zapata Sifuentes[1], César J. Huallpa[3], Gianfranco Landa Bianchi[1], Adriana Castillo Chávez[1], Romina Garavito-Salini Casas[1], Guillermo Uceda-Campos[4] and Roberto Pineda Chavarria[1]*

[1]Facultad de Ciencias Biológicas, Universidad Ricardo Palma, Lima, Peru, [2]Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, Brazil, [3]Facultad de Ciencias, Universidad Nacional Agraria La Molina, Lima, Peru, [4]Facultad de Ciencias Biológicas, Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Peru

OPEN ACCESS

Edited by:
*Basil Britto Xavier,
University of Antwerp, Belgium*

Reviewed by:
*Ritesh Tandon,
University of Mississippi,
United States
Ludmila Chistoserdova,
University of Washington,
United States*

***Correspondence:**
*Santiago Justo Arevalo
santiago.jus.are@usp.br;
santiago.justo@urp.edu.pe*

Specialty section:
*This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology*

Received: *30 September 2020*
Accepted: *25 January 2021*
Published: *17 February 2021*

Citation:
Justo Arevalo S, Zapata D, Huallpa CJ, Landa G, Castillo Chávez A, Garavito-Salini R, Uceda-Campos G and Pineda RC (2021) Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes Normalized by COVID-19 Cases During the Pandemic. Front. Microbiol. 12:612432. doi: 10.3389/fmicb.2021.612432

Since the identification of SARS-CoV-2, a large number of genomes have been sequenced with unprecedented speed around the world. This marks a unique opportunity to analyze virus spreading and evolution in a worldwide context. Currently, there is not a useful haplotype description to help to track important and globally scattered mutations. Also, differences in the number of sequenced genomes between countries and/or months make it difficult to identify the emergence of haplotypes in regions where few genomes are sequenced but a large number of cases are reported. We propose an approach based on the normalization by COVID-19 cases of relative frequencies of mutations using all the available data to identify major haplotypes. Furthermore, we can use a similar normalization approach to tracking the temporal and geographic distribution of haplotypes in the world. Using 171,461 genomes, we identify five major haplotypes or operational taxonomic units (OTUs) based on nine high-frequency mutations. OTU_3 characterized by mutations R203K and G204R is currently the most frequent haplotype circulating in four of the six continents analyzed (South America, North America, Europe, Asia, Africa, and Oceania). On the other hand, during almost all months analyzed, OTU_5 characterized by the mutation T85I in nsp2 is the most frequent in North America. Recently (since September), OTU_2 has been established as the most frequent in Europe. OTU_1, the ancestor haplotype, is near to extinction showed by its low number of isolations since May. Also, we analyzed whether age, gender, or patient status is more related to a specific OTU. We did not find OTU's preference for any age group, gender, or patient status. Finally, we discuss structural and functional hypotheses in the most frequently identified mutations, none of those mutations show a clear effect on the transmissibility or pathogenicity.

Keywords: SARS-CoV-2, COVID-19, viral pandemic, phylogenomic, global analysis, epidemiology, haplotypes, operational taxonomic units

INTRODUCTION

COVID-19 was declared a pandemic by the World Health Organization on March 11th, 2020 (Cucinotta and Vanelli, 2020), with around 71 million cases and 1.6 million deaths around the world (December 14th, 2020; WHO, 2020), quickly becoming the most important health concern in the world. Several efforts to produce vaccines, drugs, and diagnostic tests to help in the fight against SARS-CoV-2 are being mounted in a large number of laboratories all around the world.

Since the publication on January 24th, 2020 of the first complete genome sequence of SARS-CoV-2 from China (Zhu et al., 2020), thousands of genomes have been sequenced in a great number of countries on all six continents and were made available in several databases. This marks a milestone in scientific history and gives us an unprecedented opportunity to study how a specific virus evolves in a worldwide context. As of November 30th, 2020, the global initiative on sharing all influenza data (GISAID) database (Shu and McCauley, 2017) contained 171,461 genomes with at least 29,000 sequenced bases.

Several analyses have been performed to identify SARS-CoV-2 variants around the world, most of them on a particular group of genomes using limited datasets (For example, Castillo et al., 2020; Franco-Muñoz et al., 2020; Maitra et al., 2020; Saha et al., 2020). In March 2020, two major lineages were proposed based on position 8,782 and 28,144 using a data set of 103 genomes (Tang et al., 2020) which was followed by a particularly interesting proposal that identified the same major lineages (named A and B) and other sublineages (Rambaut et al., 2020).

To complement these current classification systems, we consider that haplotypes description and nomenclature could help to better track important mutations that are currently circulating in the world. Identification of SARS-CoV-2 haplotypes aids in understanding the evolution of the virus and may improve our efforts to control the disease.

To perform a reasonable analysis of the worldwide temporal and geographical distribution of SARS-CoV-2 haplotypes, we need to take into account the differences in the number of sequenced genomes in months and countries or continents. Thus, we first used a data set of 171,461 complete genomes to estimate the worldwide relative frequency of nucleotides in each SARS-CoV-2 genomic position and found nine mutations with respect to the reference genome EPI_ISL_402125 with normalized relative frequencies (NRFp) representing to be present in more than 9,500,000 COVID-19 cases. After that, using a total of 109,953 complete genomes without ambiguous nucleotides from GISAID, we performed a phylogenetic analysis and correlated the major branches with SARS-CoV-2 variants which can be classified into five haplotypes or operational taxonomic units (OTUs) based on the distribution of the nine identified nucleotide positions in our NRFp analysis. After that, we analyzed the geographical and temporal worldwide distribution of OTUs normalized by the number of COVID-19 cases. Also, we attempt to correlate these OTUs with patient status, age, and gender information. Finally, we discuss the current hypothesis of the most frequent mutations on protein structure and function. All this information will be continuously updated in our publicly available web-page.[1]

MATERIALS AND METHODS

Normalized Frequency Analysis of Each Base or Gap by Genomic Position

To perform the mutation frequency analysis, we first downloaded a total of 171,461 complete and high coverage genomes from the GISAID database (as of November 30th, 2020). This set of genomes was aligned using ViralMSA using default parameter settings, and EPI_ISL_402125 SARS-CoV-2 genome from nt 203 to nt 29,674 as the reference sequence (Li, 2018; Moshiri, 2020). Subalignments corresponding to genomes divided by continent-month combinations were extracted and relative frequencies of each base or gap in each genomic position were calculated $\left(RF_{p,m-c}\right)$ using a python script. These relative frequencies were multiplied by the number of cases reported in the respective continent-month combination $\left(CN_{m-c}\right)$ obtaining an estimation of the number of cases that present a virus with a specific base or gap in a specific genomic position $\left(RF_pCN_{m-c}\right)$. Finally, we added the RF_pCN_{m-c} of each subalignment and divided it by the total number of cases in the world $\left(\sum_{m-c} RF_pCN_{(m-c)_i}\right)/TCN_w$. This procedure allows us to obtain a relative frequency normalized by cases of each base or gap in each genomic position $\left(NRF_p\right)$. The number of cases of each country was obtained from the European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. We used the number of cases of countries with at least one genome sequenced and deposited in GISAID database. Also, we just consider in the analysis month-continent combinations with at least 90 genomes sequenced.

Phylogenetic Tree Construction

Using an alignment of the 109,953 complete, high coverage genomes without ambiguities, we estimated a maximum likelihood tree with Fasttree v2.1.10 with the next parameters: -nt -gtr -gamma -sprlength 1000 -spr 10 -refresh 0.8 -topm 1.5 close 0.75 (Price et al., 2009, 2010), after the generation of the tree, we improved topology using -boot 1000 and the first output tree as an input using -intree option. To generate the rooted tree (against EPI_ISL_402125), we used the R package treeio, and to generate tree figures with continent or date information by tip, we used the ggtree package in R (Yu et al., 2017; Yu, 2020).

OTUs Determination

Mutations respect to EPI_ISL_402125 with NRFp greater than 0.18 were extracted from the alignment of the non-ambiguous data set of 109,953 genomes and were associated with the

[1] http://sarscov2haplofinder.urp.edu.pe/

whole-genome rooted tree using the MSA function from the ggtree package (Yu et al., 2017; Yu, 2020) in R. Then, we visually examined to identify the major haplotypes based on these positions, designated as OTUs. Haplotypes identification based on our NRFp calculation reduced the bias of the different number of genomes sequenced in each continent and each month by integrating the less biased information of the number of cases. Although, other biases are more difficult, if possible, to reduce or eliminate.

Analysis of OTUs Geographical Distribution

In this analysis, we randomly separate the genomes into six samples of 28,576 genomes each. Genomes in each sample were divided by continents and by months. In these divisions, OTUs relative frequencies were calculated for each OTU in each month-continent combination $\left(O_nF_{m-c}\right)$. Then, we multiplied these $\left(O_nF_{m-c}\right)$ frequencies by the number of cases corresponding to the respective month-continent $\left(CN_{m-c}\right)$ to obtain an estimation of the number of cases caused by a specific OTU in a respective month-continent $\left(O_nCN_{m-c}\right)$. After, these products were grouped by continents, and those from the same continent were added and then divided by the total number of cases in the continent analyzed $\left(\sum_{m-c1} O_nCN_{m-c_i}\right)/TCN_{c_i}$. Thus, obtaining a frequency normalized by cases for each OTU in each continent. Finally, following this procedure in each sample, we statistically compared the mean of those six samples using the package "ggpubr" in R with the non-parametric Kruskal-Wallis test, and pairwise statistical differences were calculated using non-parametric Wilcoxon test from the same R package. The number of cases of each country was obtained from the European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. We used the number of cases of countries with at least one genome sequenced and deposited in GISAID database. Also, we just consider in the analysis month-continent combinations with at least 90 genomes sequenced.

Analysis of OTUs Temporal Distribution

Following a similar procedure used in the geographical analysis, we now grouped the products O_nCN_{m-c} by months, added them, and then divided by the total number of cases in the analyzed month $\left(\sum_{m_i-c} O_nCN_{m_i-c}\right)/TCN_{m_i}$. As in the geographical analysis, the mean of the six samples was statistically compared using the same procedures and with exactly the same considerations of month-continent combinations.

Analysis of Age, Gender, and Patient Status With OTUs Distribution

We determine if OTUs have a preference for age or gender, or cause a COVID-19 with a specific severity. For patient status and age information, we selected populations with at least 45 genomes in the category to analyze and at least two times the total number of genomes (for example, Asia – February has 58 asymptomatic genomes and 613 total genomes). For the gender analysis, we selected sample populations with at least 250 genomes in the category to analyze and at least two times the total number of genomes (for example, USA – March has 2,079 genomes from female patients and 9,287 genomes with or without gender information). In each selected sample, we used the total data (all genomes corresponding to that continent-month combination) and the data with category information (for example, male, female, asymptomatic, severe, 16–30 years, etc.). We randomly divided these two groups of genomes into three samples and calculated OTUs frequencies. The mean of the frequency of each OTU was compared between the two groups using the non-parametric Wilcoxon or Kruskal-Wallis statistical test. In the case of age information, the relative frequencies of each OTU of the total genomes and the genomes with category information were correlated using Spearman correlation. All plots were produced in R using "ggpubr" and ggplot2.

RESULTS AND DISCUSSION

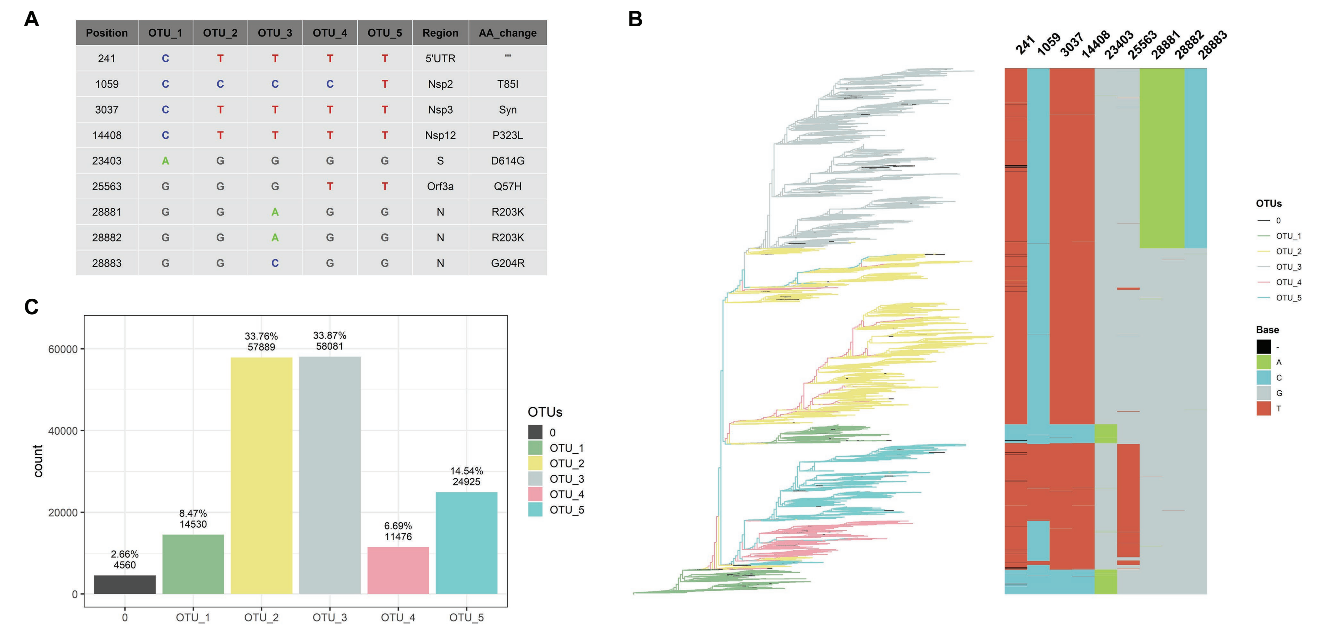
Mutations Frequency Analysis

The GISAID database contains 171,461 genomes with at least 29,000 sequenced bases; from these, 109,953 genomes do not present ambiguities (as of November 30th). With an alignment of the 171,461 genomes, we performed a normalized relative frequency analysis of each nucleotide in each genomic position (NRFp; see Materials and Methods section for details). This normalization was performed to detect relevant mutations that could appear in regions where few genomes were sequenced (**Supplementary Figure S1** shows that no correlation exists between the number of cases and the number of sequenced genomes). Using this NRFp analysis, we identified nine positions estimated to be in more than 9,500,000 COVID-19 cases (more than 0.18 NRFp; **Figure 1A** and **Supplementary Figure S2A**) plus many other mutations with NRFp between 0.00 and 0.18 (**Supplementary Figures S2B,C**).

The nine most frequent mutations (NRFp greater than 0.18) comprise seven non-synonymous mutations, one synonymous mutation, and one mutation in the 5'-UTR region of the SARS-CoV-2 genome (**Figure 1A**). The three consecutive mutations G28881A, G28882A, and G28883C falls at the 5' ends of the forward primer of "China-CDC-N" (**Supplementary Table S1**). Because these three mutations are at the 5' ends, it is unlikely that those mutations greatly affect amplification efficiency. The other six mutations do not fall within regions used by qRT-PCR diagnostic kits (**Supplementary Table S1**). All these nine mutations have been already identified in other studies (Kern et al., 2020; Korber et al., 2020; Pachetti et al., 2020; Yun, 2020), although with different frequencies mainly due to the absence of normalization.

OTUs Identification

After NRFp analysis, we estimated a maximum likelihood tree using the whole-genome alignment of the 109,953 complete



Position	OTU_1	OTU_2	OTU_3	OTU_4	OTU_5	Region	AA_change
241	C	T	T	T	T	5'UTR	""
1059	C	C	C	C	T	Nsp2	T85I
3037	C	T	T	T	T	Nsp3	Syn
14408	C	T	T	T	T	Nsp12	P323L
23403	A	G	G	G	G	S	D614G
25563	G	G	G	T	T	Orf3a	Q57H
28881	G	G	A	G	G	N	R203K
28882	G	G	A	G	G	N	R203K
28883	G	G	C	G	G	N	G204R

FIGURE 1 | Five haplotypes [or operational taxonomic units (OTUs)] based on nine positions can classify 97% of the genomes. **(A)** Table showing haplotype of each OTU, regions, and aminoacids changes caused by these mutations. **(B)** Rooted tree of 109,953 SARS-CoV-2 complete and non-ambiguous genomes associated with an alignment of nine genomic positions (241, 1,059, 3,037, 14,408, 23,403, 25,563, 28,881, 28,882, and 28,883) showing a good correlation between haplotypes (OTUs) based on these nine positions. Tips of the tree were colored based on the OTU. **(C)** Bar diagram showing OTUs distribution of the genomes (0 correspond to unclassified genomes).

genomes without ambiguities. Then, we associated the branches of the tree with an alignment of the nine positions (241, 1,059, 3,037, 14,408, 23,403, 25,563, 28,881, 28,882, and 28,883). We noted that combinations of those nine positions represent five well-defined groups in the tree (**Figure 1B**). Using these combinations, we defined five haplotypes that allow us to classify more than 97% of the analyzed genomes (**Figure 1C**), a great part of the remaining not classified genomes are due to the absence of sequencing corresponding to position 241. We named these haplotypes as OTUs.

OTU_1 was considered the ancestor haplotype due to its identity with the first isolated genomes (EPI_ISL_402125 and EPI_ISL_406801) with characteristic C241, C3037, C14408, and A23403. This OTU_1 comprised genomes with T or C in position 8,782 and C or T in 28,144. In other analyses, these mutations divide SARS-CoV-2 strains into two lineages. For instance, at the beginning of the pandemic, Tang et al. (2020) showed linkage disequilibrium between those positions and named them as S and L lineages. Rambaut et al. (2020) used these positions to discriminate between their proposed major lineages A and B. Those mutations did not reach the estimated number of 9,500,000 COVID-19 cases, indicating that a small number of these genomes emerged during the pandemic in comparison with other variations.

A SARS-CoV-2 isolated on January 25th in Australia is at present the first belonging to OTU_2 (**Supplementary Figure S3**). Showing simultaneously four mutations different to OTU_1 (C241T, C3037T, C14408T, and A23403G), OTU_2 is the first group containing the D614G and the P323L mutations in the spike and nsp12 protein, respectively. Korber et al. (2020) analyzed the temporal and geographic distribution of this mutation separating SARS-CoV-2 into two groups, those with D614 and those with G614. Tomaszewski et al. (2020) analyzed the entropy of variation of these two mutations (D614G and P323L) until May. Apparently, OTU_2 is the ancestor of two other OTUs (OTU_3 and OTU_4), as shown in the maximum likelihood tree (**Figure 1B**). OTU_2 is divided into two major branches, one that originates OTU_3 and another more recent branch characteristic from Europe (see below, worldwide geographical distribution of OTUs).

On February 16th in the United Kingdom, a SARS-CoV-2 with three adjacent mutations (G28881A, G28882A, and G28883C; **Supplementary Figure S3**) in N protein was isolated. These three mutations (together with those that characterized OTU_2) define OTU_3. The maximum likelihood tree shows that OTU_4 comes from OTU_2. OTU_4 does not present mutations in N protein; instead, it presents a variation in Orf3a (G25563T). Finally, OTU_5 presents all the mutations of OTU_4 plus one nsp2 mutation (C1059T).

These nine mutations have been separately described in other reports but, to our knowledge, they have not yet used been used together to classify SARS-CoV-2 haplotypes during the pandemic. The change of relative frequencies of those mutations analyzed individually showed that just in few cases, mutations that define haplotypes described here appear independently (**Supplementary Figure S4**). For example, the four mutations that define OTU_2 (C241T, C30307T, C14408T, and A23403G) rarely had been described separately and similarly with mutations that characterize OTU_3 (G28881A, G28882A, and G28883C; **Supplementary Figure S4**). Thus, in this case, analysis of haplotypes will be identical results that if we analyzed those mutations independently.

The fact that we were able to classify more than 97% of the complete genomes data set (**Figure 1C**) shows that, at least to the present date, this classification system covers almost all the currently known genomic information around the world. Also, most of the unclassified tips appear within a clade allowing us to easily establish their phylogenetic relationships to a haplotype. Thus, at the moment, this system can be of practical use to analyze the geographical and temporal distribution of haplotypes during these 11 months of 2020. For convenience, we presented **Supplementary Table S2** that contains the relation between our identified OTUs and their relationships with pangolin lineages (Rambaut et al., 2020) and GISAID clades (Shu and McCauley, 2017).

Worldwide Geographic Distribution of OTUs

Using our OTUs classification, we analyzed the worldwide geographic distribution during 11 months of 2020. We began by plotting continental information in the ML tree of the unambiguous complete genomes (**Figure 2A**) and observed some interesting patterns. For instance, all continents contain all OTUs; also, is relatively clear that most isolates belonging to OTU_5 come from North America (**Figure 2A**). Furthermore, the biggest branch of OTU_2 is almost exclusively filled by genomes from Europe, is interesting to note that this branch also contains genomes isolated in the last months analyzed showing its relatively recent appearance (see below, the worldwide temporal distribution of OTUs). However, this approach does not allow us to evaluate continents with less sequenced genomes (**Supplementary Figure S5A**), such as South America, Oceania, and Africa. Also, it is possible that fine differences can be found in the frequency of one OTU concerning another in each continent. These differences are not observed at this level of analysis.

To better analyze which were the most prevalent OTUs in each continent, we analyzed all the complete genomes in the GISAID database (171,461 genomes). In this analysis, we compared the mean of the frequency of OTUs normalized by cases in each continent of six randomly selected groups of genomes (see Materials and Methods section for more details).

This approach more clearly illustrates that OTU_5 was the most prevalent in North America, followed by OTU_2 and OTU_3, the least prevalent were OTU_1 and OTU_4 (**Figure 2B**). The first genomes in North America belonged to OTU_1 (**Supplementary Figure S6**). Since March, North America was dominated by OTU_5 (**Supplementary Figure S6**). OTU_5 has six of the nine high-frequency genomic variations described (all except those in N protein; **Figure 1A**).

South America presents a greater OTU_3 frequency (**Figure 2C**) that was established in April (**Supplementary Figure S5**). This observation correlates well with studies focused in South America that detect the establishment of D614G mutation at the end of March (mutation presents in OTU_2, OTU_3, OTU_4 and OTU_5) and a high frequency of pangolin lineage B1.1 in Chile and in general in South America that contains the same characteristics mutations that our OTU_3 (Castillo et al., 2020; Franco-Muñoz et al., 2020). Unfortunately, few genomes are reported in South America for September, October, and November (24 genomes in total in the three months), hindering a correct analysis of frequencies in these months. Similarly, OTU_3 was most prevalent in Asia, Oceania, and Africa (**Figures 2E–G**). With other OTUs with least than 0.3 NRFp (**Figures 2E–G**). Wu et al. (2020) report high incidence of mutations that define OTU_3 in Bangladesh, Oman, Russia, Australia, and Latvia. At the haplotype level, OTU_3 presents mutations in the N protein that apparently increases the fitness of this group in comparison with OTU_2 (OTU_2 does not present mutations in N; **Figure 1A**). Thus, four of the six continents analyzed present an estimation of more than 50% COVID-19 cases with a SARS-CoV-2 with the three mutations in the N protein. We, therefore, believe that it is important to more deeply study if exists positive fitness implications for these mutations.

Europe presents an interesting pattern (**Figure 2D**), it follows a similar pattern to South America, Asia, Oceania, and Africa until July (**Supplementary Figure S6**), with OTU_3 as the predominant. Then, in August, OTU_2 increased its frequency, and since September, OTU_2 is the most prevalent in Europe (**Figure 2D**). This could be caused by the appearance of mutations in the background of OTU_2 (such as those described in Justo et al., 2020b) with greater fitness than those of OTU_3 or due to other effects (i.e., founder effects) after the relaxation of lockdown policies.

Worldwide Temporal Distribution of OTUs

A rooted tree was estimated with the 109,953 genomes data set and labeled by date (**Figure 3A**). Here, we can observe that OTU_1 is mostly labeled with colors that correspond to the first months of the pandemic, expected due to its relation with the first genomes isolated. Clades, where OTU_2, OTU_3, OTU_4, and OTU_5 are the most prevalent, have similar distributions, with representatives mostly isolated since April. The biggest branch of OTU_2 presents a very specific temporal distribution with almost all the genomes isolated from September to November.

To gain more insight into these patterns, we estimated the most prevalent OTUs in the world during each month of the pandemic following similar steps that those done for continents (see Materials and Methods section for details). In this analysis, we did not consider December and January that present all genomes except one belonging to OTU_1 and mainly from Asia (**Supplementary Figures S6, S7**).

Analysis using the data of February from North America, Europe, and Asia showed that OTU_1 continued as the most prevalent in the world but with first isolations of OTU_2, OTU_3, OTU_4, and OTU_5 (**Figure 3B**). Analysis by continents showed that during this month, Asia and North America still had higher proportions of OTU_1, but in Europe, a more homogeneous distribution of OTU_1, OTU_2, and OTU_3 was observed (**Supplementary Figure S6**).

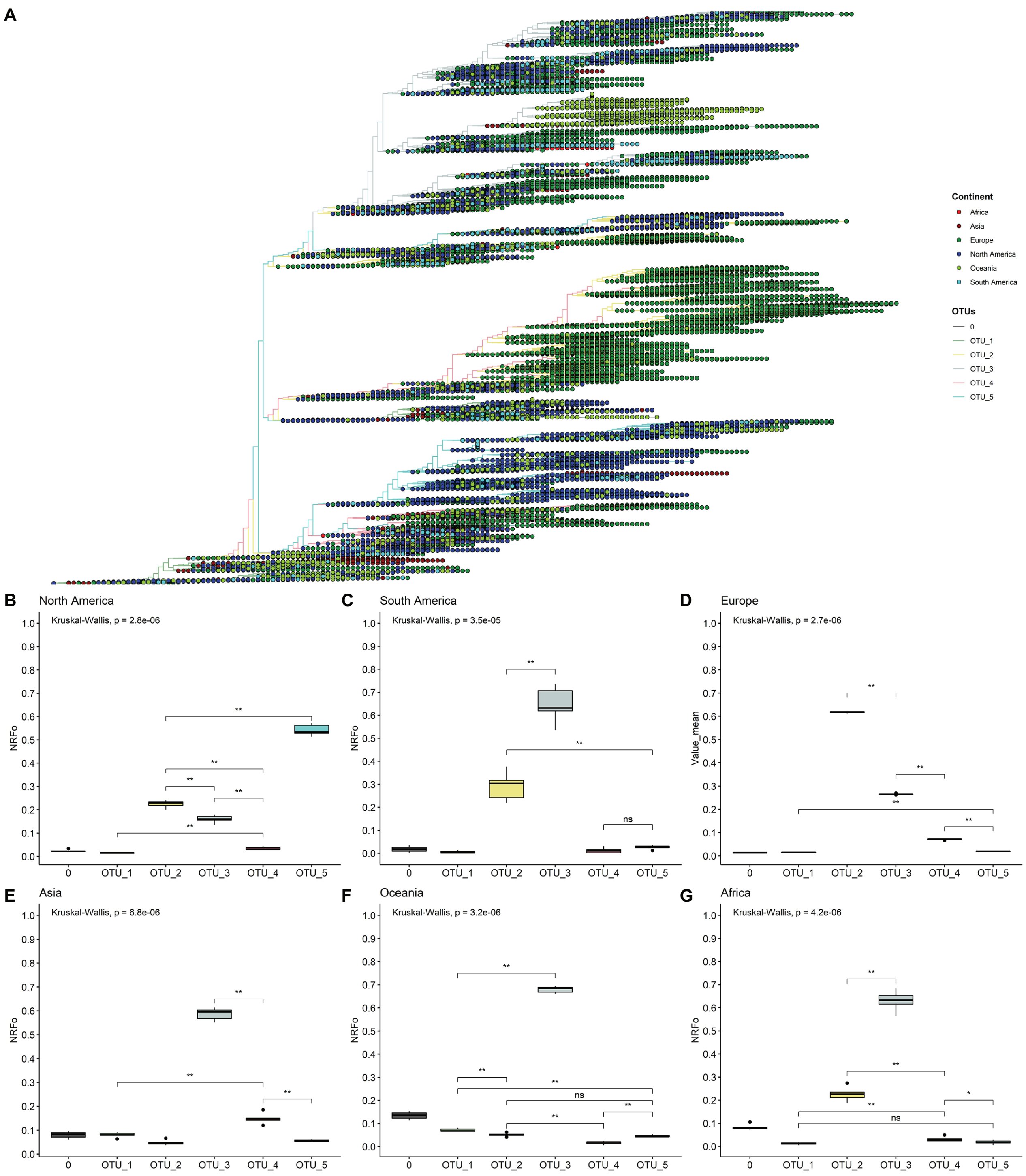


FIGURE 2 | By cases, normalized continent distribution of OTUs shows OTU_3 as the most prevalent in four of six continents. **(A)** Unrooted tree of complete non-ambiguous genomes, tips were colored according to OTUs, and points in each tip were colored according to the continent. **(B–G)** Boxplots of normalized relative frequencies of OTUs in each continent from December 2019 to November 2020 (**B**, North America; **C**, South America; **D**, Europe; **E**, Asia; **F**, Oceania; and **G**, Africa). **p* < 0.05; ***p* < 0.01.

In March, when the epicenter of the pandemic moved to Europe and North America, but cases were still appearing in Asia, OTU_2, OTU_3, and OTU_5 increased their prevalence but OTU_1 remained slightly as the most prevalent during this month (**Figure 3C**). Interestingly, OTU_4 remained in relatively low frequencies (**Figure 3C**). This month contains

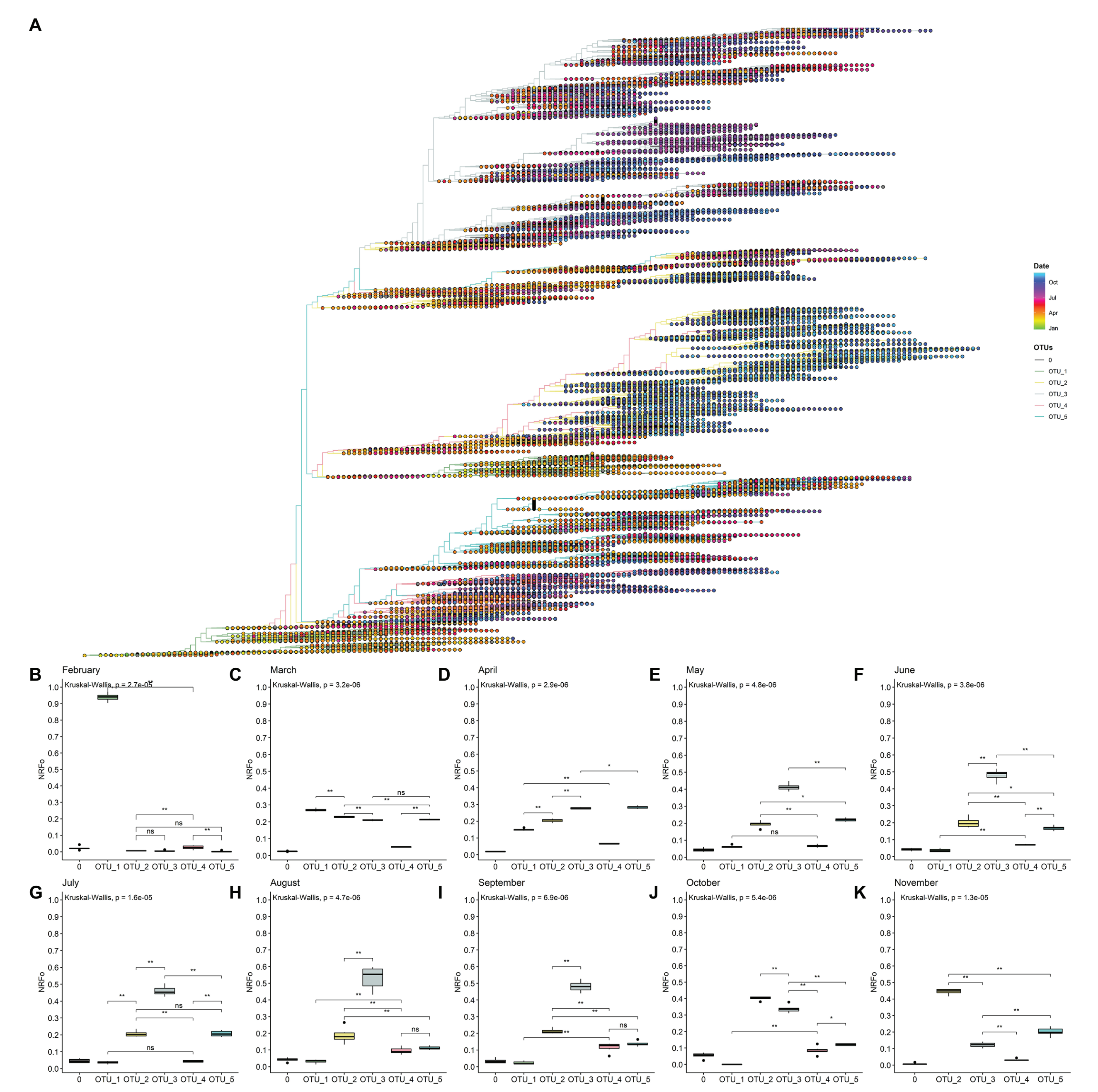


FIGURE 3 | By cases, normalized temporal distribution of OTUs showed OTU_3 as the most prevalent until September. **(A)** Rooted tree of complete non-ambiguous genomes showing temporal distribution. Tips were colored by OTUs and points in each tip were colored according to the collection date. **(B–K)** Boxplot of OTUs global distribution in each month (**B**, February; **C**, March; **D**, April; **E**, May; **F**, June; **G**, July; **H**, August; **I**, September; **J**, October; and **K**, November). *$p < 0.05$; **$p < 0.01$.

the more homogenous OTUs distribution in a worldwide context, but with some OTUs more prevalent in each continent (**Supplementary Figure S6**).

During April, OTU_1 continued its downward while OTU_3 and OTU_5 increased their presence (**Figure 3D**) probably due to its higher representation (compared to March) in several continents such as South America, North America, and Europe (**Supplementary Figure S6**). During this month, Africa showed a high prevalence of OTU_2 (**Supplementary Figure S6**). We also witnessed the establishment of OTU_3 in South America and OTU_5 in North America (**Supplementary Figure S6**).

May, June, and July showed a similar pattern, with OTU_3 as the most prevalent due to its high frequencies in South America, Oceania, and Europe (**Figures 3E–G** and **Supplementary Figure S6**). North America maintains OTU_5 as the most prevalent and Oceania showed a relatively homogenous pattern. During these months, OTU_2 had intermediate frequencies in all continents resulting in intermediate frequencies all over the world (**Figures 3E–G** and **Supplementary Figure S6**). OTU_1 and OTU_4 representatives were reported during these months but with very low frequencies.

In August and September, we detected a slightly higher frequency of OTU_4 compared to the previous months (**Figures 3H,I**) with no significant differences with OTU_5. In September in Europe, OTU_3 stopped being the most frequent. Instead, OTU_2 was the most frequent in this month in Europe (**Supplementary Figure S6**). In October and November, OTU_2 has increased its frequency rapidly (**Figures 3J,K**) mainly due to a large number of cases and reported genomes belonging to this OTU_2 in Europe in October and November. Due to the few genomes currently available in GISAID for all continents, except for Europe and North America during November, just these two continents were analyzed in the last month.

Also, it is important to mention that there are not many enough genomes reported for September, October, and November for South America, so during these months, OTUs frequencies of this continent were not considered.

Age, Gender, and Patient Status Relation With OTUs

Relating the distribution of haplotypes according to patient information can help to determine the preference of some OTUs for some characteristics of the patients. Thus, we analyze OTUs distribution according to age, gender, and patient status information available as metadata in the GISAID database.

Unfortunately, just 26.11% of the 171,461 genomes analyzed have age and gender information (**Supplementary Figure S8**). In the case of patient status information, we noted that GISAID categories are not well organized and we had to reclassify the information into three categories: Asymptomatic, Mild, and Severe (**Supplementary Figure S9A**). Using this classification scheme, we found that 99.14% (169,979 genomes) were not informative, 0.1% (175 genomes) falls in the Asymptomatic category, 0.33% (562 genomes) in the Mild category, and 0.43% (745 genomes) could be classified as Severe (**Supplementary Figure S9B**).

Using this limited data, we attempt to determine whether any OTU causes an asymptomatic, mild, or severe infection more frequently. We look for significant differences between the relative frequencies of the OTUs in total samples and samples with known patient information. If we found differences, it would mean that some OTU could be more or less related to one type of infection. Here, we analyzed just the month-continent combination with at least 45 genomes with information of one type of infection and at least two times of genomes with any information (for example, Asia – February has 58 Asymptomatic genomes and 613 total genomes). Ten combinations meet these criteria, one in the asymptomatic category, one in the mild, and eight in the severe. None of the OTUs frequencies in samples with patient status information were significative different from the frequencies in the total population of the month-continent analyzed (**Figure 4**). Thus, we concluded that none of the OTUs are related to an asymptomatic, mild, or severe COVID-19, at least in the populations analyzed.

Age information was also analyzed in the same manner. In general, although some differences were detected as significant, those were not consistently maintained between different populations analyzed (**Supplementary Figures S10A–J**). Furthermore, none difference reaches a value of p less than 0.01 (except for OTU_4 in North America). Since heterogeneity between countries information is possible, we think that these small differences are more likely due to these heterogeneities and we cannot strongly conclude that some age groups are more related to a specific OTU. Additionally, a strong positive correlation between total relative frequencies of OTUs and relative frequencies by age groups in month-continent was found, meaning that those two frequencies are similar in most of the analyzed populations (**Supplementary Figure S10K**).

A similar approach was done using gender information, but in this case, due to the greater quantity of information, we used more restrictive filter parameters. Thus, we selected country-month combinations with at least 250 genomes with male or female information and two times total genomes information (for instance USA – March has 2079 genomes from female patients and 9,287 genomes with or without gender information). Again, we did not find OTU's preference for a specific gender (**Supplementary Figure S11**).

Description of the Most Frequent Mutations

C241T

The C241T mutation is present in the 5'-UTR region. In coronaviruses, the 5'-UTR region is important for viral transcription (Madhugiri et al., 2014) and packaging (Masters, 2019). Computational analysis showed that this mutation could create a TAR DNA-binding protein 43 (TDP43) binding site (Mukherjee and Goswami, 2020), TDP43 is a well-characterized RNA-binding protein that recognizes UG-rich nucleic acids (Kuo et al., 2014) described to regulate splicing of pre-mRNA, mRNA stability and turnover, and mRNA trafficking and can also function as a transcriptional repressor and protect mRNAs under conditions of stress (Lee et al., 2011). Experimental studies are necessary to confirm different binding constants of TDP43 for the two variants of 5'-UTR and its *in vivo* effects.

C1059T

Mutation C1059T lies on Nsp2. Nsp2 does not have a clearly defined function in SARS-CoV-2 since the deletion of Nsp2 from SARS-CoV has little effect on viral titers and so maybe dispensable for viral replication (Graham et al., 2005). However, Nsp2 from SARS-CoV can interact with prohibitin 1 and 2 (PBH1 and PBH2; Cornillez-Ty et al., 2009), two proteins involved in several cellular functions including cell cycle progression (Wang et al., 1999), cell migration (Rajalingam et al., 2005), cellular differentiation (Sun et al., 2004), apoptosis (Fusaro et al., 2003), and mitochondrial biogenesis (Merkwirth and Langer, 2008).

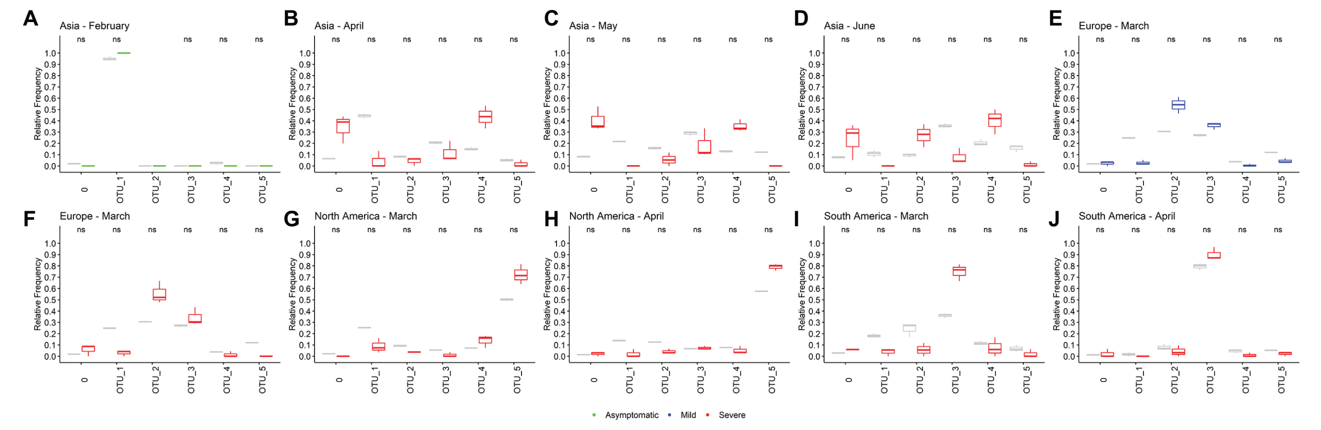


FIGURE 4 | Operational taxonomic units are not related to the COVID-19 severity. **(A–J)** Ten different sample populations were analyzed, none of the OTUs frequencies shows significative differences between the total samples and samples taken from genomes with patient status information. Boxplots showed the distribution of three samples, total frequencies are showed in gray, and frequencies from samples with patient status information are colored according the category (green, asymptomatic; blue, mild; and red, severe).

C3037T

Mutation C3037T is a synonymous mutation in Nsp3; therefore, it is more difficult to associate this change with an evolutionary advantage for the virus. This mutation occurred in the third position of a codon. One possibility is that this changes the frequency of codon usage in humans increasing expression or any other of the related effects caused by synonymous codon change (some of them reviewed in Mauro and Chapel, 2014).

C3037T causes a codon change from TTC to TTT. TTT is more frequently present in the genome of SARS-CoV-2 and other related coronaviruses compared to TTC (Gu et al., 2020) but in humans, the codon usage of TTT and TTC are similar (Mauro and Chapel, 2014). The reason why TTT is more frequent in SARS-CoV-2 is unknown but seems to be a selection related to SARS-CoV-2 and not to the host. Another option is genetic drift.

C14408T

The C14408T mutation changes P323 to leucine in Nsp12, the RNA-dependent RNA polymerase of SARS-CoV2 (**Supplementary Figures S12A,B**). P323 together with P322 ends helix 10 and generate a turn that is followed by a beta-sheet (**Supplementary Figure S12C**). Leucine at position 323 could form hydrophobic interactions with the methyl group of L324 and the aromatic ring of F396 creating a more stable variant of Nsp12 (**Supplementary Figure S12E**). In concordance with this, protein dynamics simulations showed a stability increase of the Nsp12 P323L variant (Chand and Azad, 2020). In the absence of P322, the mutation P323L would probably be disfavored due to the flexibilization of the turn at the end of helix 10. Experimental evidence is necessary to confirm these hypotheses and to evaluate their impact on protein function.

A23403G

An interesting protein to track is spike protein (**Supplementary Figure S13A**) due to its importance in SARS-CoV-2 infectivity. It has been suggested that the D614G change in the S1 domain that results from the A23403G mutation generates a more infectious virus, less spike shedding, greater incorporation in pseudovirions (Zhang et al., 2020), and higher viral load (Korber et al., 2020).

How these effects occur at the structural level remains unclear, although some hypotheses have been put forward: (1) We think that there is no evidence for hydrogen-bond between D614 and T859 mentioned by Korber et al. (2020), and distances between D614 and T859 are too long for a hydrogen bond (**Supplementary Figure S13B**), (2) distances between Q613 and T859 (**Supplementary Figure S13C**) could be reduced by increased flexibility due to D614G substitution, forming a stabilizing hydrogen bond, and (3) currently available structures do not show salt-bridges between D614 and R646 as proposed by Zhang et al. (2020; **Supplementary Figure S13D**).

G25563T

Orf3a (**Supplementary Figure S14A**) is required for efficient *in vitro* and *in vivo* replication in SARS-CoV (Castaño-Rodriguez et al., 2018). It has been implicated in inflammasome activation (Siu et al., 2019), apoptosis (Chan et al., 2009), and necrotic cell death (Yue et al., 2018) and has been observed in Golgi membranes (Padhan et al., 2007) where pH is slightly acidic (Griffiths and Simons, 1986). Kern et al. (2020) showed that Orf3a preferentially transports Ca^{+2} or K^{+} ions through a pore (**Supplementary Figure S14B**). Some constrictions were described in this pore, one of them formed by the side chain of Q57 (**Supplementary Figure S14C**).

Mutation G25563T produces the Q57H variant of Orf3a (**Supplementary Figure S14C**). It did not show significant differences in expression, stability, conductance, selectivity, or gating behavior (Kern et al., 2020). We modeled Q57H mutation and we did not observe differences in the radius of constriction (**Supplementary Figure S14C**) formed by residue 57 but we observed slight differences in the electrostatic surface due to the ionizability of the histidine side chain (**Supplementary Figure S14D**).

G28881A, G28882A, and G28883C

N protein is formed by two domains and three disordered regions. The central disordered region named LKR was shown to interact directly with RNA (Chang et al., 2009) and other proteins (Luo et al., 2005), probably through positive side chains; also, this region contains phosphorylation sites able to modulate the oligomerization of N protein (Chang et al., 2013).

Mutation G28883C that changes a glycine for arginine at position 204 contributes one more positive charge to each N protein. Mutations G28881A and G28882A produce a change from arginine to lysine. These two positive amino acids probably have a low impact on the overall electrostatic distribution of N protein. However, change from R to K could alter the probability of phosphorylation in S202 or T205. Using the program NetPhosK (Blom et al., 2004), we observed different phosphorylation potential in S202 and T205 between G28881-G28882-G28883 (RG) and A28881-A28882-C28883 (KR; **Supplementary Figure S15**). Other authors proposed that these mutations could change the molecular flexibility of N protein (Rahman et al., 2020).

CONCLUDING REMARKS

Here, we present a complete geographical and temporal worldwide distribution of SARS-CoV-2 haplotypes from December 2019 to November 2020. We identified nine high-frequency mutations. These important variations (asserted mainly by their frequencies) need to be tracked during the pandemic.

Our haplotypes description showed to be phylogenetically consistent, allowing us to easily monitor the spatial and temporal changes of these mutations in a worldwide context. This was only possible due to the unprecedented worldwide efforts in the genome sequencing of SARS-CoV-2 and the public databases that rapidly share the information.

Our geographical and temporal analysis showed that OTU_3 is currently the more frequent haplotype circulating in four of six continents (Africa, Asia, Oceania, and South America), result that is in accordance with other studies (Mercatelli and Giorgi, 2020) that showed GISAID clade GR (that corresponds to our OTU_3) as the most prevalent in the world; however, they did not report the currently predominance of OTU_2 in Europe (clade G for GISAID). Intriguingly, OTU_3 never reached frequencies higher than OTU_5 in North America. In Europe, currently and different from the tendency from May to July, OTU_2 is now much more commonly isolated than OTU_3. Why mutations R203K and G204R have such frequencies in most of the continents, why in North America, those mutations were not so successful and why currently Europe is dominated by OTU_2 are open questions. Some studies showed that at the moment, there are not mutations that significative increase the fitness of the SARS-CoV-2 (Kepler et al., 2020; van Dorp et al., 2020).

Although OTU_1 was the only and the most abundant haplotype at the beginning of the pandemic, now its isolation is rare. This result shows an expected adaptation process of SARS-CoV-2. This enunciate does not mean that SARS-CoV-2 is now more infectious or more transmissible.

In the next months, these haplotypes description will need to be updated, identification of new haplotypes could be performed by combining the identification of new frequent mutations and phylogenetic inference. We will continue monitoring the emergence of mutations that exceed our proposed cut-off of 0.18 NRFp and this information will be rapidly shared with the scientific community through our web page.[2] This will also be accompanied by a continuous update of haplotypes information. During the peer-review process of this manuscript, we identify several other mutations near to the cut-off proposed that were reported in Justo et al. (2020b).

Using information of specific populations, we showed no preference for patient's features (age, gender, or type of infection) by OTUs. Thus, mutations that define those haplotypes do not have a relevant impact on the severity of the disease neither are implied preferentially in infections to males, females, or age.

Finally, although more studies need to be performed to increase our knowledge of the biology of SARS-CoV-2, we were able to make hypotheses about the possible effects of the most frequent mutations identified. This will help in the development of new studies that will impact vaccine development, diagnostic test creation, among others.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: gisaid.org.

AUTHOR CONTRIBUTIONS

SA, DS, CH, GB, AC, and RG-SC conceived, initiated, and coordinated the project. SA performed the phylogenetic analyses, geographical and temporal analyses. GU-C wrote python scripts used in data processing and analyses. SA, DS, CH, GB, AC, RG-SC and RC performed the structural analysis. The manuscript was written by SA, DS, CH, and GB. All authors discussed the methodologies and results, and read and approved the manuscript.

FUNDING

We thank to the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) graduate scolarship (to SA; 2015/13318-4) and Universidad Ricardo Palma (URP) for APC financing.

[2] http://sarscov2haplofinder.urp.edu.pe/

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://doi.org/10.1101/2020.07.12.199414 (Justo et al., 2020a).

We are very grateful to the GISAID Initiative and all its data contributors, i.e., the authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared *via* the GISAID Initiative, on which this research is based. Complete acknowledgments of the 171,461 genomes used are available in **Supplementary Material (SF1–SF20)**.

We thank Shaker Chuck Farah (Institute of Chemistry – University of Sao Paulo) for English writing corrections and helpful comments. Also, we thank Aline Maria da Silva (Institute of Chemistry – University of Sao Paulo), Joao Renato Rebello Pinho (Albert Einstein Hospital – Sao Paulo) and Deyvid Amgarten (Albert Einstein Hospital – Sao Paulo) for its helpful comments. To the Ricardo Palma University High-Performance Computational Cluster (URPHPC) managers Gustavo Adolfo Abarca Valdiviezo and Roxana Paola Mier Hermoza at the Ricardo Palma Informatic Department (OFICIC) for their contribution in programs and remote use configuration of URPHPC. To Gladys Arevalo Chong for her figure style suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.612432/full#supplementary-material

REFERENCES

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the aminoacid sequence. *Proteomics* 4, 1633–1649. doi: 10.1002/pmic.200300771

Castaño-Rodriguez, C., Honrubia, J., Gutierrez-Alvarez, J., DeDiego, M., Nieto-Torres, J., Jimenez-Guardeño, J., et al. (2018). Role of severe acute respiratory syndrome coronavirus viroporins E, 3a, and 8a in replication and pathogenesis. *mBio* 9, e02325–e02417. doi: 10.1128/mBio.02325-17

Castillo, A. E., Parra, B., Tapia, P., Lagos, J., Arata, L., Acevedo, A., et al. (2020). Geographical distribution of genetic variants and lineages of SARS-CoV-2 in Chile. *Front. Public Health* 8:562615. doi: 10.3389/fpubh.2020.562615

Chan, C., Tsoi, H., Chan, W., Zhai, S., Wong, C., Yao, X., et al. (2009). The ion channel activity of the SARS-coronavirus 3a protein is linked to its pro-apoptotic function. *Int. J. Biochem. Cell Biol.* 41, 2232–2239. doi: 10.1016/j.biocel.2009.04.019

Chand, G., and Azad, G. (2020). Identification of novel mutations in RNA-dependent RNA ploymerases of SARS-CoV-2 and their implications. bioRxiv [Preprint]. doi: 10.1101/2020.05.05.079939

Chang, C., Chen, C., Chiang, M., Hsu, Y., and Huang, T. (2013). Transient oligomerization of the SARS-CoV N protein – Implication for virus ribonucleoprotein packaging. *PLoS One* 8:e65045. doi: 10.1371/journal.pone.0065045

Chang, C., Hsu, Y., Chang, Y., Chao, F., Wu, M., Huang, Y., et al. (2009). Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein implications for ribonucleocapsid protein packaging. *J. Virol.* 83, 2255–2264. doi: 10.1128/JVI.02001-08

Cornillez-Ty, C., Liao, L., Yates, J., Kuhn, P., and Buchmeier, M. (2009). Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J. Virol.* 83, 10314–10318. doi: 10.1128/JVI.00842-09

Cucinotta, D., and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Biomed.* 91, 157–160. doi: 10.23750/abm.v91i1.9397

Franco-Muñoz, C., Álvarez-Díaz, D., Laiton-Donato, K., Wiesner, M., Escandón, P., Usme-Ciro, J., et al. (2020). Substitutions in spike and nucleocapsid proteins of SARS-CoV-2 circulating in South America. *Infect. Genet. Evol.* 85:104557. doi: 10.1016/j.meegid.2020.104557

Fusaro, G., Dasgupta, P., Rastogi, S., Joshi, B., and Chellappan, S. (2003). Prohibitin induces the transcriptional activity of p53 and is exported from the nucleus upon apoptotic signaling. *J. Biol. Chem.* 278, 47853–47861. doi: 10.1074/jbc.M305171200

Graham, R., Sims, A., Brockway, S., Baric, S., and Denison, M. (2005). The nsp2 replicase protein of murine hepatitis virus and severe acute respiratory syndrome coronavirus is dispensable for viral replication. *J. Virol.* 79, 13399–13411. doi: 10.1128/JVI.79.21.13399-13411.2005

Griffiths, G., and Simons, K. (1986). The trans Golgi network: sorting at the exit site of the golgi complex. *Science* 234, 438–443. doi: 10.1126/science.2945253

Gu, H., Chu, D., Peiris, M., and Poon, L. (2020). Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. bioRxiv [Preprint]. doi: 10.1101/2020.02.15.950568

Justo, S., Zapata, D., Huallpa, C., Landa, G., Castillo, A., Garavito-Salini, R., et al. (2020a). Global geographic and temporal analysis of SARS-CoV-2 haplotypes normalized by COVID-19 cases during the pandemic. bioRxiv [Preprint]. doi: 10.1101/2020.07.12.199414

Justo, S., Zapata, D., Huallpa, C., Landa, G., Castillo, A., Garavito-Salini, R., et al. (2020b). Analysis of the dynamics and distribution of SARS-CoV-2 mutations and its possible structural and functional implications. bioRxiv [Preprint]. doi: 10.1101/2020.11.13.381228

Kepler, L., Hamins-Puertolas, M., and Rasmussen, D. (2020). Decomposing the sources of SARS-CoV-2 fitness variation in the United States. bioRxiv [Preprint]. doi: 10.1101/2020.12.14.422739

Kern, D., Sorum, B., Hoel, C., Sridharan, S., Remis, J., Toso, D., et al. (2020). Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. bioRxiv [Preprint]. doi: 10.1101/2020.06.17.156554

Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812.e19–827.e19. doi: 10.1016/j.cell.2020.06.043

Kuo, P., Chiang, C., Wang, Y., Doudeva, L., and Yuan, H. (2014). The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids. *Nucleic Acids Res.* 42, 4712–4722. doi: 10.1093/nar/gkt1407

Lee, E., Lee, V., and Trojanowski, J. (2011). Gains or losses: molecular mechanisms of TDP43-mediated neurodegeneration. *Nat. Rev. Neurosci.* 13, 38–50. doi: 10.1038/nrn3121

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Luo, H., Chen, Q., Chen, J., Chen, K., Shen, X., and Jiang, H. (2005). The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. *FEBS Lett.* 579, 2623–2628. doi: 10.1016/j.febslet.2005.03.080

Madhugiri, R., Fricke, M., Marz, M., and Ziebuhr, J. (2014). RNA structure analysis of alphacoronavirus terminal genome regions. *Virus Res.* 194, 76–89. doi: 10.1016/j.virusres.2014.10.001

Maitra, A., Chawla, M., Raheja, H., Biswas, N., Chakraborti, S., Kumar, A. M., et al. (2020). Mutations in SARS-CoV-2 viral RNA identified in Eastern India: possible implication for the ongoing outbreak in India and impact on viral structure and host susceptibility. *J. Biosci.* 45:76. doi: 10.1007/s12038-020-00046-1

Masters, P. (2019). Coronavirus genomic RNA packaging. *Virology* 537, 198–207. doi: 10.1016/j.virol.2019.08.031

Mauro, V., and Chapel, S. (2014). A critical analysis of codon optimization in human therapeutics. *Trends Mol. Med.* 20, 604–613. doi: 10.1016/j.molmed.2014.09.003

Mercatelli, D., and Giorgi, F. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. *Front. Microbiol.* 11:1800. doi: 10.3389/fmicb.2020.01800

Merkwirth, C., and Langer, T. (2008). Prohibitin function within mitochondria: essential roles for cell proliferation and cristae morphogenesis. *Biochim. Biophys. Acta* 1793, 27–32. doi: 10.1016/j.bbamcr.2008.05.013

Moshiri, N. (2020). ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* btaa743. doi: 10.1093/bioinformatics/btaa743 [Epub ahead of print]

Mukherjee, M., and Goswami, S. (2020). Global cataloguing of variations in untranslated regions of viral genome and prediction of key host RNA binding protein-microRNA interactions modulating genome stability in SARS-CoV-2. bioRxiv [Preprint]. doi: 10.1101/2020.06.09.134585

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18:179. doi: 10.1186/s12967-020-02344-6

Padhan, K., Tanwar, C., Hussain, A., Hui, P., Lee, M., Cheung, C., et al. (2007). Severe acute respiratory syndrome coronavirus Orf3a protein interacts with caveolin. *J. Gen. Virol.* 88, 3067–3077. doi: 10.1099/vir.0.82856-0

Price, M., Dehal, P., and Arkin, A. (2009). FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Price, M., Dehal, P., and Arkin, A. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rahman, M., Islam, M., Alam, A., Islam, I., Hoque, M., Akter, S., et al. (2020). Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *J. Med. Virol.* 1–19. doi: 10.1002/jmv.26626 [Epub ahead of print]

Rajalingam, K., Wunder, C., Brinkmann, V., Churin, Y., Hekman, M., Sievers, C., et al. (2005). Prohibitin is required for RAS-induced RAF-MEK-ERK activation and epithelial cell migration. *Nat. Cell Biol.* 7, 837–843. doi: 10.1038/ncb1283

Rambaut, A., Holmes, E., Hill, V., O'Toole, A., Hill, V., McCrone, J., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407. doi: 10.1038/s41564-020-0770-5

Saha, O., Hossain, M., and Rahaman, M. (2020). Genomic exploration light on multiple origin with potential parsimony-informative sites of the severe acute respiratory syndrome coronavirus 2 in Bangladesh. *Gene Rep.* 21:100951. doi: 10.1016/j.genrep.2020.100951

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* 22, 1–3. doi: 10.2807/1560-7917.ES.2017.22.13.30494

Siu, K., Yuen, K., Castaño-Rodriguez, C., Ye, Z., Yeung, M., Fung, S., et al. (2019). Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J.* 33, 8865–8877. doi: 10.1096/fj.201802418R

Sun, L., Liu, L., Yang, X., and Wu, Z. (2004). Akt binds prohibitin 2 and relieves its repression of MyoD and muscle differentiation. *J. Cell Sci.* 117, 3021–3029. doi: 10.1242/jcs.01142

Tang, X., Wi, C., Li, X., Song, Y., Yao, X., Wu, X., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* 7, 1012–1023. doi: 10.1093/nsr/nwaa036

Tomaszewski, T., DeVries, R., Dong, M., Bhatia, G., Norsworthy, M., Zheng, X., et al. (2020). New pathways of mutational change in SARS-CoV-2 proteomes involve regions of intrinsic disorder important of virus replication and release. *Evol. Bioinform.* 16, 1–18. doi: 10.1177/1176934320965149

van Dorp, L., Richard, D., Tan, C., Shaw, L., Acman, M., and Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* 11:5986. doi: 10.1038/s41467-020-19818-2

Wang, S., Nath, N., Adlam, M., and Chellappan, S. (1999). Prohibitin, a potential tumor suppressor, interacts with RB and regulates E2F function. *Oncogene* 18, 3501–3510. doi: 10.1038/sj.onc.1202684

World Health Organization (2020). Available at: https://covid19.who.int/ (Accessed August 25, 2020).

Wu, S., Tian, C., Liu, P., Guo, D., Zheng, W., Huang, X., et al. (2020). Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. *J. Med. Virol.* doi: 10.1002/jmv.26597 [Epub ahead of print]

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* 69, 1–18. doi: 10.1002/cpbi.96

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Yue, Y., Nabar, N., Shi, C., Kamenyeva, O., Xiao, X., Hwang, I., et al. (2018). SARS-Coronavirus open reading frame-3a drives multimodal necrotic cell death. *Cell Death Dis.* 9, 1–15. doi: 10.1038/s41419-018-0917-y

Yun, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implication. *Genomics* 112, 3588–3596. doi: 10.1016/j.ygeno.2020.04.016

Zhang, L., Jackson, C., Mou, H., Ojha, A., Rangarajan, E., Izard, T., et al. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv [Preprint]. doi: 10.1101/2020.06.12.148726

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix 4: Curriculum Vitae.

CURRICULUM VITAE

SANTIAGO JUSTO ARÉVALO
AGE: 30
CIVIL STATUS: SINGLE
DATE OF BIRTH: JANUARY 7 OF 1992
NATIONALITY: PERUVIAN (DNI: 46700251)
ADDRESS IN PERU: AV. MALECÓN ASENT. H. NUEVO LURÍN MZ. A5 LT. 01 LIMA
Phone number: (+55) 11 94 532 4772
Professional address in Brazil: Instituto de Química - USP, Av. Prof. Lineu Prestes 748, Cidade Universitária, São Paulo, SP, Brazil 05508-000
Email: santiago.jus.are@usp.br, santiago.justo@urp.edu.pe, sanjusare_712@hotmail.com

EDUCATION

Sao Paulo University – Institute of Chemistry – Department of Biochemistry
PhD. Program in Biochemistry (In course)
Project: Structural Engineering of the Active Site of the GGDEF Domains to Produce New Second Messengers **2015-2021**

Ricardo Palma University – Faculty of Biological Sciences
Undergraduate Program in Biology (completed) **2010-2014**

ACADEMIC ACTIVITIES

A) PUBLICATIONS AND ARTICLES

Genomic characterization of Bacillus strains capable to degrade cyanide isolated from mine tailings in Peru, with emphasis on Bacillus safensis and its cyanide-degrading enzyme CynD
Justo S, Zapata D, Cuba A, Brescia M, Monge C, Farage L, Marques P, Morais C, Guerra A, Quiñones M, Farah C, Setubal J, da Silva A.
Applied and Environmental Microbiology. 2022.
In review.

Updating the phylodynamics of Yellow Fever Virus 2016–2019 Brazilian outbreak with new 2018 and 2019 São Paulo genomes
Salles A, Nastri A, Ho Y, Casadio L, Amgarten D, **Justo S**, Gomes-Gouvea M, Carrilho F, de Mello F, Rebello J.
Frontiers in Microbiology. 2022.
https://doi.org/10.3389/fmicb.2022.811318

Biodegradation of cyanide using a Bacillus subtilis strain isolated from artisanal gold mining tailings
Alvarez C, Vallenas A, **Justo S**, Romano D, Soares J.
Brazilian Journal of Chemical Engineering. 2022.
https://doi.org/10.1007/s43153-022-00228-4

Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures
Justo S, Zapata D, Huallpa C, Landa G, Castillo A, Garavito-Salini R, Uribe C, Uceda-Campos G, Pineda R.
Scientific Reports. 2021.
https://doi.org/10.1038/s41598-021-97267-7

Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes Normalized by COVID-19 Cases during the Pandemic
Justo S, Zapata D, Huallpa C, Landa G, Castillo A, Garavito-Salini R, Uceda-Campos G, Pineda R.
Frontiers in Microbiology. 2021.
https://doi.org/10.3389/fmicb.2021.612432

Analysis of the Dynamics and Distribution of SARS-CoV-2 Mutations and its Possible Structural and Functional Implications
Justo S, Zapata D, Huallpa C, Landa G, Castillo A, Garavito-Salini R, Uceda-Campos G, Pineda R.

BioRxiv. 2020.
https://doi.org/10.1101/2020.11.13.381228

Bioleaching of metal from waste stream using a native strain of Acidithiobacillus isolated from a coal mine drainage
Kazue S, **Justo S**, Alvarez C, Quiñones M, Soares J, Romano D
The Canadian Journal of Chemical Engineering. 2019.
https://doi.org/10.1002/cjce.23519

Efectos del extracto de β -cariofileno de Piper nigrum en la hiperactivación espermática y en la capacidad fecundante deTtetrapygus niger (molina, 1782) ''erizo negro de mar''
Zapata D, Wixsan J, Hinostroza K, **Justo S**, Gonzales-Figueroa H.
Biotempo. 2019.
https://doi.org/10.31381/biotempo.v16i2.2529

A bipartite periplasmic receptor-diguanylate cyclase pair (XAC2383-XAC2382) in the bacterium Xanthomonas citri.
Texeira R, Guzzo C, **Justo S**, Andrade M, Abrahao J, de Souza R, Farah C.
Journal of Biological Chemistry. 2018.
https://doi.org/10.1074/jbc.RA118.003475

Molecular Dynamics and In Silico Analysis of Oligomerization Surfaces of CYND Enzymes
Médico A, Bustamante C, Pineda R, **Justo S.**
Proceedings of MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th edition.
https://doi.org/10.3390/mol2net-04-05126

Modeling and Analysis of AG:IGE Interface of House Dust Mite Allergens of Group 1
Churasacari T, Zapata D, Maza J, Arica A, Pineda R, **Justo S**.
Proceedings of MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th edition
https://doi.org/10.3390/mol2net-04-05125

Isolation and characterization of cyanide degrading bacteria from artisanal gold mine tailings.
Monge C, Rosario C, Vallenas A, **Justo S**, Tenorio J.
The 32nd International Conference on solid Waste Technology and Management, 2017. pp.641-648

Scientific Journal of the Microbiology Lab of Biological Sciences Faculty Ricardo Palma University. Volume 2 - Nº1 - 2016
General Editor (Santiago Justo)
Articles in this journal:

- **Isolation and identification of *Salmonella enterica* from guinea pig with "Salmonelosis" signs.** Pag. 14-17.
- **Isolation of a specific phage against *Salmonella* Typhimurium isolated from intestine sample of *Cavia porcellus*** Pag. 18-21.
- **The cyclic dinucleotides, its roles and the ubiquitous protein domain GGDEF.** Pag. 22-30

Scientific Journal of the Microbiology Lab of Biological Sciences Faculty Ricardo Palma University. Volume 1 - Nº1 and Nº2 - 2014
General Editor (Santiago Justo)
Articles in these Journals:

- **Marine Bacteria: "Best Answers Comes in Little Bottles".** Vol. 1 Nº1: pag. 10
- **Determination of Bacterial Genera on the Marine Environment of Cantolao Beach – La Punta -Callao.** Vol. 1 Nº2: pag. 20-23
- **Antimicrobial activity against clinically important bacterial strains of an Acetone-Water Extract of** *Macroscystis pyrifera* **(C. Agardh 1820)** Vol. 1 Nº2: pag. 15-19
- **Antibacterial Effect of the Pyocianin of** *Pseudomonas aeruginosa* **against** *E. coli* **and** *S. aureus* Vol. 1 Nº2: pag. 24-27
- **General aspects of the Guinea Pig breeding and its infectious agent: *Salmonella*** Vol. 1 Nº2: pag. 28-33

Origins of the springs of Costa Verde beach in Lima Perú (Origen de los manantiales de la Costa Verde)
Rojas R, Montoya M, Mamani E, Maguiña J, Montoya E, Baltuano O, Bedregal P, Coria L, Guerra A, **Justo S,** Churasacari T
Cornell University Library **arXiv:1305.2158**
Encuentro Científico Internacional Vol 9. Nº 2 Marzo 2013 ISSN 1813 - 0194 – Lima – Perú

B) UNDERGRADUATE RESEARCH PROJECTS

Functional studies on the biogenesis of type IV pilus in *Xanthomonas axonopodis* pv. citri
Department of Biochemistry – Institute of Chemistry – São Paulo University
Principal Investigators: PhD. German Dunger and Prof. Dr. Chuck S. Farah.

Structural Studies on the Active site of GGDEF domains
Department of Biochemistry – Institute of Chemistry – São Paulo University
Principal Investigator: Prof. Dr. Chuck S. Farah

Development of a Recombinant Vaccine to Protects Guinea pigs against *Salmonella*
Department of Microbiology and Immunology - Faculty of Biological Sciences – Ricardo Palma University
Principal Investigator: Prof. Alcides Guerra Santa Cruz

Cloning, Expression and Purification of a Recombinant Allergen Der f 6
Department of Microbiology and Immunology - Faculty of Biological Sciences – Ricardo Palma University
Principal Investigator: Prof. Alcides Guerra Santa Cruz

C) PARTICIPATION IN SCIENTIFIC MEETINGS

Distribución global geográfica y temporal de haplotipos de SARS-CoV-2 normalizado por casos de COVID-19
International Scientific Meeting Summer 2021 – Lima – Perú **2021**

Global geographic and temporal analysis of SARS-CoV-2 Haplotypes normalized by COVID-19 cases
COVID-19 Dynamics and Evolution – San Diego – USA **2020**

Distribución global geográfica y temporal de haplotipos de SARS-CoV-2 normalizado por casos de COVID-1
Bioinformática aplicada a la investigación de COVID-19 – Lima – Perú **2020**

Análisis temporal y geográfico mundial de haplotipos de SARS-CoV-2 muestra patrones de distribución diferenciados
International Scientific Meeting Winter 2020 – Lima – Perú **2020**

New linear and cyclic dinucleotides formation by wild-type GGDEF domains
30 Congresso Brasileiro de Microbiologia 2019 – Alagoas – Brazil **2019**

Genome sequencing of three Bacillus strains isolated in Peru and search of genes related to cyanide metabolism
30 Congresso Brasileiro de Microbiologia 2019 – Alagoas – Brazil **2019**

Herramientas bioinformáticas para el estudio de proteínas
Charla CIBH: La bioteconología y sus aplicaciones – Lima – Perú. **2019**

Estudios en la degradación bacteriana de cianuro. Un enfoque bioinformático, estructural y funcional
International Scientific Meeting Summer 2019 – Lima – Perú **2019**

Estudios en la degradación bacteriana de cianuro. Un enfoque bioinformático, estructural y funcional
I Encuentro Internacional de Investigadores en Ciencia e Ingenieria – Lima – Perú **2018**

Enzymatic production of new cyclic dinucleotides by wild type GGDEF domains
Xantomonadaceae (Xantho) meeting – Sao Paulo – Brazil **2018**

A fast and effective methodology for purification of cyanide-degrading enzyme from B. pumilus
International Congress of Biotechnology and Innovation (ICBi 2018) - Lima – Perú **2018**

Phylogenetic and sequence-based analysis of prokaryotic cyanide dihydratases (CynDs)
International Congress of Biotechnology and Innovation (ICBi 2018) - Lima – Perú **2018**

A low molecular weight protein induced as a function of cyanide concentration in B. licheniformis.

International Congress of Biotechnology and Innovation (ICBi 2018) - Lima – Perú **2018**

In silico analysis of 334 mutation in SLC17A5 explain defects in lysosomal transport of sialic acid.
International Congress of Biotechnology and Innovation (ICBi 2018) - Lima – Perú **2018**

Studies on the microbial degradation of cyanide. A functional, structural and bioinformatic approach.
International Congress of Biotechnology and Innovation (ICBi 2018) - Lima – Perú **2018**

Bioinformática estructural: Una forma de predecir el comportamiento de proteínas para crear soluciones a problemas.
International Scientific Meeting Summer 2018 – Lima – Perú **2018**

Structural bioinformatics: A from to predict protein behavior to create solution to problems.
III Jornada internacional en Biociencias (Hamutay 2017) - Lima – Perú **2017**

Molecular dynamics and in silico analysis of Cyanide dihydratases (CynD) of Bacillus pumilus
Pseudomonas stutzeri
III Jornada internacional en Biociencias (Hamutay 2017) - Lima – Perú **2017**

Modeling and analysis of the hydrogen bond interaction network between 5H8 antibody
and house dust mite allergens of group 1
III Jornada internacional en Biociencias (Hamutay 2017) - Lima – Perú **2017**

Modeling and Analysis of Ag:IgE interface of house dust mite allergens of group 1
Second international Conference in Bioinformatics, Simulations and Modeling (iCBSM 2107) – Talca – Chile **2017**

Molecular dynamics and in silico analysis of oligomerization surface of CynD enzymes
Second international Conference in Bioinformatics, Simulations and Modeling (iCBSM 2107) – Talca – Chile **2017**

Molecular dynamics and in silico analysis of flexible regions of CynD from Pseudomonas stutzeri
Interuniversity meeting – Expobiol 2017 – Lima – Perú **2017**

Molecular dynamics and in silico analysis of oligomerization surface of CynD from Bacillus pumilus
Interuniversity meeting – Expobiol 2017 – Lima – Perú **2017**

Modeling and structural analysis of Ag:Ac interactions of group 1 allergens of house dust mites
Interuniversity meeting – Expobiol 2017 – Lima – Perú **2017**

Isolation and Characterization of Cyanide degrading Bacteria From Artisanal Gold Mine Tailings.
The 32nd International Conference on Solid Waste Technology and Management **2017**

Salmonelosis en los cuyes y Contaminacion por Cianuro. ¿Cómo el estudio estructural de proteínas puede
ayudar a resolver estos problemas?
International Scientific Meeting Summer 2017 – Lima – Perú **2017**

Structural engineering of the active site of diguanylate cyclase (GGDEF) domains to produce new second messengers.
Latin American Protein Society meeting (V LAPS) – Rio de Janeiro – Brazil **2016**

Development of a subunit vaccine against Salmonella for Cavia porcellus "Cuy" based in the surface
Protein PilA (PpdD)
Interuniversity meeting – Expobiol 2016 – Lima – Perú **2016**

Isolation and Characterization of Cyanide degrading bacteria from mining tailings
Interuniversity meeting – Expobiol 2016 – Lima – Perú **2016**

¿Cómo la Ingeniería Estructural de Proteínas nos Ayuda a Resolver Problemas?
International Scientific Meeting Summer 2016 – Lima – Perú **2016**

Influence of Xac0258, Xac0259 and FimX in Xac Type IV pilus-dependant motility
Xantomonadaceae (Xantho) meeting – Sao Paulo – Brazil **2015**

Determination of Bacterial Genera on the Marine Environment of Cantolao Beach – La Punta -Callao.
IV Congress of Marine Sciences of Perú – IV CONCIMAR – Lima – Perú **2014**

Antimicrobial activity against clinically important bacterial strains of an Acetone-Water Extract of
Macroscystis pyrifera (C. Agardh 1820)
International Scientific Meeting of Winter 2014 and XV CONEBIOL – Lima – Perú **2014**

Determination of Bacterial Genera on the Marine Environment of Cantolao Beach – La Punta –Calla
International Scientific Meeting of Winter 2014 and XV CONEBIOL – Lima – Perú **2014**

Antibacterial Effect of the Pyocianin of Pseudomonas aeruginosa against E. coli and S. aureus
International Scientific Meeting of Winter 2014 and XV CONEBIOL – Lima – Perú **2014**

Standardization of egg extraction protocol, massive culture and observation of early development of Panagrellus redivivus
XIV National Congress of Biological Students – XIV CONEBIOL – Iquitos – Perú **2013**

Isolation and Characterization of Salmonella enterica lipopolysaccharides
IV Interuniversity meeting - Expobiol 2012 – Lima – Perú **2013**

Antibiotic effect of pyocianin and pyoverdin pigments of three strains of Pseudomonas aeruginosa
IV Interuniversity meeting - Expobiol 2012 - Lima – Perú **2013**

Estimate of the percentages of more frequent solid waste types in Arica Beach, Lurín, Lima.
III Congress of Marine Sciences of Perú – III CONCIMAR – Lima – Perú **2012**

Evaluation of the content of saponins of 13 varieties of Chenopodium quinoa W. (Amaranthaceae) cultured in Perú
XIII National Congress of Biological Students – XII CONEBIOL – Ica – Perú
2012

Evaluation of the contamination of the fresh water springs of La Estrella beach – Miraflores - Perú
IV Interuniversity meeting - Expobiol 2012 – Lima – Perú **2012**

D) ORGANIZER OF SCIENTIFIC EVENTS

I International course of Crystallography of Macromolecules, Lima - Peru
Organizer – Expositor **2019**

Introduction to the Structural study of proteins, Lima – Peru
Organizer – Expositor **2016**

Workshop about Basic Bioinformatic Tools – Ricardo Palma University
Organizer – Expositor **2015**

Workshop about Scientific publication – Ricardo Palma University
Organizer of the Event **2014**

Conferences Cycle about Cancer and Immunity – Ricardo Palma University
Organizer of the Event **2014**

XV National Congress of Biological Students – XVCONEBIOL – Lima – Perú
Organizer – Vice-president **2014**

Worshop about Bacterial Transformation using Calcium Chloride – XVCONEBIOL
Organizer - Expositor **2014**

Workshop with Undergraduate students on the EUROAMERICANO High-School – Lima – Perú
Organizer - Expositor
DNA electrophoresis, ELISA assays, Gram coloration. **2014**

Conferences Cycle of Microbial Biotechnology in the National Development – Ricardo Palma University
Organizer of the event **2013**

Activities in benefit of the Community in Santiago de Surco District – Ricardo Palma University
Organizer of Environmental Campaigns in the city. **2012**

I National Congress of Quality and Food Safety – CONCIA – Ricardo Palma University
Organizer as part of the Scientific comission **2012**

E) AWARDS AND DISTINCTIONS

Winner of the 1st Prize in the Oral Presentation Contest during International Congress of Biotechnology

and Innovation.
ICBi 2018 – Lima – Peru
Studies on the microbial degradation of cyanide. A functional, structural and bioinformatic approach **2018**

Selected to do an internship on the Oxford Protein Production Facility (OPPF)
Oxford – United Kingdom
High-Throughput cloning, expression, purification and crystallization experiments with Xac0610 protein **2018**

Winner of the 1st Prize in the Oral Presentation Contest during Interuniversity meeting
Expobiol 2017 – Lima – Peru
Molecular dynamics and in silico analysis of oligomerization surface of CynD from Bacillus pumilus **2017**

Winner of the 1st Prize in the Poster Session Contest during Interuniversity meeting
Expobiol 2017 – Lima – Peru
Modeling and structural analysis of Ag:Ac interactions of group 1 allergens of house dust mites **2017**

Winner of the 2nd Prize in the Poster Session Contest during Interuniversity meeting
Expobiol 2017 – Lima – Peru
Molecular dynamics and in silico analysis of flexible regions of CynD from Pseudomonas stutzeri **2017**

Winner of a Financial Scolarship by FAPESP in PhD. Program in Sao Paulo University
Sao Paulo University – Institute of Chemistry – Department of Biochemistry
Project: Structural Engineering of the Active Site of the GGDEF Domains to Produce New Second Messengers **2015**

Winner of a Grant to develop an Undergraduate Thesis Project - III Contest of Undergraduate Thesis
Ricardo Palma - University
Project: "Development of a Recombinant Vaccine for the control of Salmonella outbreak
in *Cavia porcellus* "Guinea Pig" **2015**

Diploma of Honor for the 1st Position in the Biological Sciences Faculty
Ricardo Palma University **2014**

Winner of the 1st Prize in the Poster Session Contest during the XV National Congress of Biology students
XV CONEBIOL – Lima – Peru
Antibacterial Effect of the Pyocianin of Pseudomonas aeruginosa against E. coli and S. aureus **2014**

Winner of the 1st Prize in the Oral Presentation Contest on the XV National Congress of Biology students
XV CONEBIOL – Lima – Peru
Antimicrobial activity against clinically important bacterial strains of an Acetone-Water Extract of Macroscystis pyrifera (C. Agardh 1820) **2014**

Winner of the 3rd Prize in the Panels Contest on the XIII National Congress of Biology students
XIII CONEBIOL – Ica – Peru
Evaluation of the content of saponins of 13 varieties of Chenopodium quinoa W. (Amarantheaceae) Cultured in Perú
2014

TEACHING EXPERIENCE

Ricardo Palma University – Biological Sciences Faculty
Laboratory assistant in Undergraduate Microbiology Course
Prof. Dr. Alcides Guerra Santa Cruz **2014**

Ricardo Palma University – Biological Sciences Faculty
Laboratory assistant in Undergraduate Microbiology Course
Prof. Dr. Alcides Guerra Santa Cruz **2013**

Ricardo Palma University – Biological Sciences Faculty
Laboratory assistant in Undergraduate Biostatistics Course
Prof. Mg. Cesar Puicon Montero **2012**

Ricardo Palma University – Biological Sciences Faculty
Laboratory assistant in Undergraduate Organic Chemistry Course
Prof. Dr. Fred García Alayo **2011**

LANGUAGES

Spanish: Advance
English: Advance

Portuguese: Advance
Chinese: Basic

REFERENCES

Prof. Dr. Shaker C. Farah
Department of Biochemistry – Insitute of Chemistry – University of Sao Paulo
Av. Prof Lineu Prestes 748
Tlf: (+55 11) 3091-3326
E-mail: chsfarah@iq.usp.br

Prof. Dr. Richard C. Garratt
Department of Physics and Informatics – Institute of Physics at São Carlos – University of Sao Paulo
Av. Trabalhador Sao-carlense, 400 Centro
Tel: (+55 16) 33739874
E-mail: Richard@ifsc.usp.br

Prof. Dr. Ray Owens
Oxford Protein Production Facility – Research Complex at Harwell – Oxford University
R92 Rutherford Appleton Laboratory, Harwell Oxford. OX11 0FA
Tel: 01235 567727
E-mail: ray.owens@strubi.ox.ac.uk

Prof. Dr. Modesto Montoya Zavaleta
Peruvian Institute of Nuclear Energy
Av. Canadá 1470 – Lima – Perú
Tel: (511) 226-0030
E-mail: modesto_montoya@yahoo.com

Prof. Dr. Mauro Quiñones Aguilar
Department of Plants and environmental Biotechnology – Faculty of Biological Sciences – Ricardo Palma University
Av. Benavides 5440 – Lima – Perú
Tel: (511) 617 6200
E-mail: mauro.quinones@gmail.com

Prof. Alcides Guerra Santa Cruz
Department of Microbiology – Faculty of Biological Sciences – Ricardo Palma University
Av. Benavides 5440 – Lima – Perú
Tel: (511) 617 6200 anexo 1429
E-mail: alcides.guerras@urp.pe